COMPARING COMPLEX NETWORKS THROUGH GRAPHLET DISTRIBUTIONS

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF SCIENCE AND ENGINEERING

2022

Miguel Eduardo Pinto da Silva

Department of Computer Science

Contents

Abstract **23**

Declaration **25**

Copyright **27**

Acknowledgements **29**

Acronyms **31**

1 Introduction **33**

1.1 Thesis contributions 36

1.2 Thesis structure 37

2 Background **39**

2.1 Network Terminology 39

2.2 Network Comparison 42

2.2.1 NetEmd 43

2.2.2 GCD 44

2.2.3 GDA 45

2.2.4 Applications of network comparison 45

2.3 Counting graphlets 46

2.3.1 ORCA 46

2.3.2 G-Tries 48

2.4 Mathematical modelling of infectious diseases 49

2.4.1 Compartmental based models 50

2.4.2 Agent-based models 55

3 Directed Network Comparison **57**
3.1 Introduction . . . 57
3.2 NetEmd with Directed Networks . . . 59
3.3 NetEmd with Dimension Reduction . . . 61
3.3.1 Motivation . . . 61
3.3.2 Principal Component Analysis . . . 63
3.3.3 Independent Component Analysis . . . 64
3.4 Experimental Setup . . . 65
3.4.1 Measure of Cluster Performance . . . 65
3.4.2 Synthetic datasets . . . 66
3.4.3 Real world datasets . . . 68
3.5 Results . . . 69
3.5.1 Undirected Results . . . 69
3.5.2 Directed Results . . . 70
3.5.3 Performance by number of ICA components . . . 72
3.5.4 Discussion on the time complexity of NetEmd . . . 75
3.5.5 Discussion on orbits from smaller graphlet sizes . . . 76
3.6 Conclusion . . . 78

4 Vaccination Discussions in Mumsnet **79**
4.1 Introduction . . . 79
4.2 Background . . . 83
4.3 Methods . . . 83
4.3.1 Data Collection . . . 83
4.3.2 Data Segmentation by Time . . . 84
4.3.3 Network Creation . . . 85
4.3.4 Network Comparison . . . 86
4.3.5 Sentiment Labelling . . . 87
4.3.6 Discordance Index . . . 89
4.4 Results . . . 90
4.4.1 Data Granularity . . . 90
4.4.2 Temporal Network Comparison . . . 97
4.4.3 Sentiment Analysis . . . 103
4.5 Discussion . . . 118
4.5.1 Summary of results . . . 118
4.5.2 The “so-what" factor . . . 121

4.5.3 Limitations . . . 121
4.6 Conclusion . . . 123

5 Contact Tracing and Testing for COVID-19 125
5.1 Introduction . . . 125
5.2 Methods . . . 128
5.2.1 Model Overview . . . 128
5.2.2 Contact Networks . . . 129
5.2.3 Interventions . . . 130
5.2.4 Outcomes . . . 133
5.2.5 Summary of changes to Covasim . . . 133
5.3 Results . . . 134
5.3.1 Asymptomatic testing and reducing number of contacts . . . 134
5.3.2 Symptomatic testing . . . 138
5.3.3 Comparing PCR and LFD tests . . . 140
5.4 Discussion . . . 144
5.5 Conclusion . . . 147

6 Conclusion 149
6.1 Summary of results . . . 149
6.2 Discussion . . . 151
6.3 Future work . . . 153

A NetEmd with different metrics 179
A.1 Area Under Precision-Recall (AUPR) . . . 179
A.1.1 Undirected Results . . . 180
A.1.2 Directed Results . . . 180
A.2 Adjusted Rand Index . . . 181
A.2.1 Undirected Results . . . 182
A.2.2 Directed Results . . . 183

B NetEmd Results tables 185

C Examples of Mumsnet threads 195

D Mumsnet - supporting figures 203

E Mumsnet - alternative data granularity **215**
E.1 Vaccination forum 215
E.2 COVID-19 forum 217
E.3 Discussion 226

F Covasim model parameters **229**

G Additional Covasim results **233**

Word Count: 46630

List of Tables

3.1 Summary statistics for the Onnela et al. and real world directed networks datasets. 68

3.2 Results for Task 1, Task 2 and the Onnela et al. [118] datasets in undirected networks. The metric used for Task 1 is the mean (and standard error of the mean) of the 16 values for $\overline{P}$ in each combination of number of nodes with network density. The metric used for Task 2 and Onnela et al. dataset is the sole value of $\overline{P}$ after comparing the 1280 and 151 networks, respectively. The parameter column indicates: graphlet size used in original NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 5); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 5; the number of orbits used by Graphlet Correlation Degree (GCD); graphlet sizes used in Graphlet Degree Distribution Agreement (GDA). The bolded values are the maximum for each task. Italic values are within one standard error of the maximum performance. Note that results for GDA with graphlet size 5 took longer than a week to return results for Task 2, at which point we stopped the computation. 69

3.3 Results for Task 2 in directed networks, for each level of reciprocity, and for the real world networks dataset. The metric used for this task is the sole value of $\overline{P}$ after comparing the 1280 and 1231 networks, respectively. The parameter column indicates: graphlet size used in directed and weighted NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 4); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 4; the number of orbits used by Directed Graphlet Correlation Degree (DGCD); graphlet sizes used in GDA. The bolded values are the maximum for each dataset. Note that results for GDA with graphlet size 4 took longer than a week to return results for Task 2 and in the real world networks dataset, at which point we stopped the computation. 71

4.1 Summary statistics for the datasets of networks from the vaccination and COVID-19 datasets. 97

4.2 Proportion of posts of positive, neutral or negative sentiment in threads of positive, neutral or negative sentiment. The table is read as: posts of sentiment A make up X% of the posts in threads of sentiment B. For example, posts of positive sentiment make up 22.5% of the posts in threads of negative sentiment or posts of negative sentiment make up 6.3% of the posts in threads of positive sentiment. 116

B.1 Results for Task 1 in directed networks, for each level of reciprocity. The metric used for evaluation is the mean (and standard error of the mean) of the 16 values for $\overline{P}$, in each combination of number of nodes, network density and reciprocity. The parameter column indicates: graphlet size used in directed and weighted NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 4); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 4; the number of orbits used by DGCD; graphlet sizes used in GDA. The bolded values are the maximum for each reciprocity level. Italic values are within one standard error of the maximum performance. 186

B.2 Results for *PCA_NetEmd* with different combinations of graphlet size and percentage of variance explained in Task 1, Task 2 and the Onnela et al. dataset in the undirected case. The metric used for Task 1 is the mean (and standard error of the mean) of the 16 values for $\overline{P}$ in each combination of number of nodes with network density. The metric used for Task 2 is the sole value of $\overline{P}$ after comparing the 1280 networks. The bolded values are the maximum in each task. Italic values are within one standard error of the maximum performance. 187

B.3 Results for *PCA_NetEmd* with different combinations of graphlet size and percentage of variance explained in Task 1 of directed networks, for each level of reciprocity. The metric used for evaluation is the mean (and standard error of the mean) of the 16 values for $\overline{P}$, in each combination of number of nodes, network density and reciprocity. The bolded values are the maximum for each reciprocity level. Italic values are within one standard error of the maximum performance. 188

B.4 Results for *PCA_NetEmd* with different combinations of graphlet size and percentage of variance explained in Task 2 in directed networks, for each level of reciprocity, and for the real world networks dataset. The metric used for evaluation is the mean (and standard error of the mean) of the 16 values for $\overline{P}$, in each combination of number of nodes, network density and reciprocity. The bolded values are the maximum for each reciprocity level. 188

B.5 Results for Task 1, Task 2 and the Onnela et al. [118] datasets in undirected networks. The metric used for Task 1 is the mean (and standard error of the mean) of the 16 values for area under precision-recall curve (AUPR) in each combination of number of nodes with network density. The metric used for Task 2 and Onnela et al. dataset is the sole value of AUPR after comparing the 1280 and 151 networks, respectively. The parameter column indicates: graphlet size used in original NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 5); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 5; the number of orbits used by GCD; graphlet sizes used in GDA. The bolded values are the maximum for each task. Italic values are within one standard error of the maximum performance. Note that results for GDA with graphlet size 5 took longer than a week to return results for Task 2, at which point we stopped the computation. . 189

B.6 Results for Task 1 in directed networks, for each level of reciprocity. The metric used for evaluation is the mean (and standard error of the mean) of the 16 values for AUPR, in each combination of number of nodes, network density and reciprocity. The parameter column indicates: graphlet size used in directed and weighted NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 4); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 4; the number of orbits used by DGCD; graphlet sizes used in GDA. The bolded values are the maximum for each reciprocity level. Italic values are within one standard error of the maximum performance. 190

B.7 Results for Task 2 in directed networks, for each level of reciprocity, and for the real world networks dataset. The metric used for this task is the sole value of AUPR after comparing the 1280 and 1231 networks, respectively. The parameter column indicates: graphlet size used in directed and weighted NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 4); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 4; the number of orbits used by DGCD; graphlet sizes used in GDA. The bolded values are the maximum for each dataset. Note that results for GDA with graphlet size 4 took longer than a week to return results for Task 2 and in the real world networks dataset, at which point we stopped the computation. 191

B.8 Results for Task 1, Task 2 and the Onnela et al. [118] datasets in undirected networks. The metric used for Task 1 is the mean (and standard error of the mean) of the 16 values for adjusted Rand index (ARI) in each combination of number of nodes with network density. The metric used for Task 2 and Onnela et al. dataset is the sole value of ARI after comparing the 1280 and 151 networks, respectively. The parameter column indicates: graphlet size used in original NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 5); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 5; the number of orbits used by GCD; graphlet sizes used in GDA. The bolded values are the maximum for each task. Italic values are within one standard error of the maximum performance. Note that results for GDA with graphlet size 5 took longer than a week to return results for Task 2, at which point we stopped the computation. . 192

B.9 Results for Task 1 in directed networks, for each level of reciprocity. The metric used for evaluation is the mean (and standard error of the mean) of the 16 values for ARI, in each combination of number of nodes, network density and reciprocity. The parameter column indicates: graphlet size used in directed and weighted NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 4); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 4; the number of orbits used by DGCD; graphlet sizes used in GDA. The bolded values are the maximum for each reciprocity level. . 193

B.10 Results for Task 2 in directed networks, for each level of reciprocity, and for the real world networks dataset. The metric used for this task is the sole value of ARI after comparing the 1280 and 1231 networks, respectively. The parameter column indicates: graphlet size used in directed and weighted NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 4); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 4; the number of orbits used by DGCD; graphlet sizes used in GDA. The bolded values are the maximum for each dataset. Note that results for GDA with graphlet size 4 took longer than a week to return results for Task 2 and in the real world networks dataset, at which point we stopped the computation. 194

E.1 Summary statistics for the datasets of networks from the vaccination and COVID-19 datasets with alternative choices of window span and jump. The values shown for the vaccination dataset correspond to a span of 6 months and jump of 2 months. For the COVID-19 dataset, the values of these two parameters are 2 weeks and 1 week, respectively. 216

F.1 Parameters modified from Covasim defaults. 230

F.2 Range of values for parameters varied in the asymptomatic testing experiment. 231

F.3 Range of values for parameters varied in the symptomatic testing experiment. 232

List of Figures

2.1 Undirected graphlets of size 2, 3 and 4 and directed graphlets of size 2 and 3. Nodes are numbered according to their orbit, counted by type (directed or undirected) from the size 2 graphlets, and nodes with the same number represent the same orbit. The background colour and shape of node boundary further emphasizes nodes belonging to the same orbit in the same graphlet. 42

2.2 Figure adapted from Hočevar and Demšar [65], showing the relationship between undirected orbit 10 and 13 from Figure 2.1. In Figure 2.2a, we see the graphlet being extended, containing nodes x, y and z, whose edges are marked by a solid line. Each of w_i is a candidate node for the extension as they are the nodes in the neighbourhood of both y and z, their edges to these nodes exist by definition and are represented through a dashed line. Finally, the dotted line is optional. Its absence leads to a four node graphlet isomorphic to G_6, as pictured in Figure 2.2b. On the other hand, its existence leads to a four node graphlet isomorphic to G_7, as pictured in Figure 2.2c. 47

2.3 Figure taken from Aparício et al. [7]. Shows the subgraph census (counting the frequency of each subgraph) using a specific G-Trie. Nodes in red mean that no candidate was found for that particular graphlet and the dotted arrows represent the search path through the G-Trie. 49

3.1 Example to illustrate the usefulness of only using orbits that appear in at least one of the networks being compared. Graphs G and H are the example provided by Aparício et al. [7]; we adapt them to directed networks G' and H', keeping the same number of orbits in each network. We measure $NetEmd(G',H') = 5 \times 10^{-5}$ and $Weighted\_NetEmd(G',H') = 0.12$. 60

3.2 Undirected networks - Task 1 73
3.3 Undirected networks - Task 2 73
3.4 Directed networks - Task 1 73
3.5 Directed networks - Task 2 74
3.6 Onnela et al. dataset 74
3.7 Real world directed networks dataset 74

4.1 User active times: distribution of how long each user remains active in each forum, obtained by taking the time difference between the date of their first and last posts. 92
4.2 Thread active times: distribution of how long each thread remains active in each forum, obtained by taking the time difference between the date of their first and last posts. 93
4.3 Number of users, threads and posts for different combinations of time jump and time span to define the time granularity in the vaccine dataset. 94
4.4 Number of users, threads and posts for different combinations of time jump and time span to define the time granularity in the COVID-19 dataset. 96
4.5 Heat map of network distances between each pair of networks in the Mumsnet vaccine dataset. Distances are calculated using *PCA_NetEmd* with size 4 graphlets and 90% explained variance. 99
4.6 Heat map of network distances between each pair of networks in the Mumsnet COVID-19 dataset. Distances are calculated using *PCA_NetEmd* with size 4 graphlets and 90% explained variance. 101
4.7 Number and proportion of threads of each sentiment label, for each data slice in the vaccination dataset. 105
4.8 Number and proportion of posts in threads of each sentiment label, for each data slice in the vaccination dataset. 106
4.9 Number and proportion of users in threads of all combinations of sentiment label, for each data slice in the vaccination dataset. 107
4.10 Number and proportion of threads of each sentiment label, for each data slice in the COVID-19 dataset. 111
4.11 Number and proportion of posts in threads of each sentiment label, for each data slice in the COVID-19 dataset. 112
4.12 Number and proportion of users in threads of each sentiment label, for each data slice in the COVID-19 dataset. 113

4.13 Percentage of posts of each sentiment label, inferred from the proportion of posts of each sentiment in threads of each sentiment in the labeled sample of the COVID-19 dataset, results shown for each data slice. 117

4.14 Average discordance index per thread in the sampled subset of the COVID-19 dataset over time, calculated as the average discordance using windows with 2 to 5 posts. 118

4.15 Average discordance index per thread in the COVID-19 dataset over time, calculated as the average discordance using windows with 2 to 5 posts. Values are inferred from the sample of sentiment labeled posts. 119

5.1 Comparison of asymptomatic testing strategies against reduction of contacts for different growth rates and contact patterns. The x-axis represents the size of the regular tester cohort for the simulations with asymptomatic testing or the percentage of contacts removed from non-household layers for the simulations with contact reduction. The y-axis shows the total number of infections, averaged over 100 runs. 135

5.2 Impact on person-days of isolation of increasing adherence to the different asymptomatic testing strategies, for different growth rates and contact patterns. The x-axis represents the size of the regular tester cohort for the simulations with asymptomatic testing or the percentage of contacts removed from non-household layers for the simulations with contact reduction. The y-axis shows the mean number of person days of isolation per population members over the 180 days of simulation, averaged over 100 runs. These include days in isolation waiting for test results (amongst true positives and true negatives); isolation days amongst those testing positive who were true positive; and isolation days among those testing positive who were false positives. The colour of each marker indicates the percentage of days spent in isolation while infected. 136

5.3 Impact of increasing the uptake on isolation on symptoms, daily symptomatic testing and Polymerase Chain Reaction (PCR) testing after being traced as a contact of a positive case. Numbers shown are the percentage decrease in the total number of infections, compared to 0% uptake on these interventions, for 0.025 and 0.3 growth rates in POLYMOD contact pattern. The top line on the x axis represents the uptake to testing after contact tracing, the bottom line represents the daily probability to testing when symptomatic. The y axis represents uptake to isolation on COVID or Influenza-like Illnesses (ILI) symptoms. . . 139

5.4 Comparison of Lateral Flow Device (LFD) against PCR tests for symptomatic or contact traced individuals, for different values of asymptomatic testing uptake and different growth rates, with contact patterns obtained from POLYMOD. The y-axis shows the total number of infections. The x-axis represents the size of the regular asymptomatic tester cohort. Markers shown represent the median of 100 simulations and the shaded area represents the interquartile range. Figure 5.4a corresponds to the twice weekly testing policy and Figure 5.4b to the testing every two days policy. 141

C.1 Organization of the Mumsnet forum, it is divided in broad categories of topics that are expanded into multiple subcategories, each with their own forum. 196

C.2 Examples of threads within the COVID-19 subforum of Mumsnet. In this case, two threads are included in our analysis, the first and last on the list. 197

C.3 Example of the organization of a thread within Mumsnet. The *original post* is coloured green and the remaining posts are shown sequentially according to the time they were posted in. 198

C.4 Examples of controversial threads in the vaccination forum of Mumsnet that may be connected to the network structure differences we observe in the time frame when these threads were posted. 199

C.5 Examples of discussions about vaccination in the COVID-19 forum of Mumsnet with representative topics that may be connected to the network structure differences we observe in the time frame when these threads were posted. 200

C.5 Continued - Examples of discussions about vaccination in the COVID-19 forum of Mumsnet with representative topics that may be connected to the network structure differences we observe in the time frame when these threads were posted. 201

D.1 Heat map of network distances between all networks in both the Mumsnet vaccination and COVID-19 dataset. Distances are calculated using *PCA_NetEmd* with size 4 graphlets and 90% explained variance. . 204

D.2 Heat map of network distances between each pair of networks in the Mumsnet vaccination dataset, for each orbit in graphlets of size up to 4. Distances are calculated using *PCA_NetEmd* with size 4 graphlets and 90% explained variance. 205

D.3 Heat map of network distances between each pair of networks in the Mumsnet COVID-19 dataset, for each orbit in graphlets of size up to 4. Distances are calculated using *PCA_NetEmd* with size 4 graphlets and 90% explained variance. 206

D.4 Heat map of network distances between all networks in both the Mumsnet vaccination and COVID-19 dataset, for each orbit in graphlets of size up to 4. Distances are calculated using *PCA_NetEmd* with size 4 graphlets and 90% explained variance. 207

D.5 Vaccination subforum - threads of each sentiment. 208

D.6 Vaccination subforum - posts in threads of each sentiment. 209

D.7 Vaccination subforum - users in threads of each sentiment. 210

D.8 COVID-19 subforum - threads of each sentiment. 211

D.9 COVID-19 subforum - posts in threads of each sentiment. 212

D.10 COVID-19 subforum - users in threads of each sentiment. 213

D.11 COVID-19 subforum - posts of each sentiment, inferred from the sampled subset of data. 214

E.1 Heat map of network distances between each pair of networks in the Mumsnet vaccination dataset, using an alternative combination of window span and jump (6 months and 2 months). Distances are calculated using *PCA_NetEmd* with size 4 graphlets and 90% explained variance. 217

E.2 Vaccination forum - threads of each sentiment - alternative data granularity. 218

E.3 Vaccination forum - posts in threads of each sentiment - alternative data granularity. 219

E.4 Vaccination forum - users in threads of each sentiment - alternative data granularity. 220

E.5 Heat map of network distances between each pair of networks in the Mumsnet COVID-19 dataset, using an alternative combination of window span and jump (2 weeks and 1 week). Distances are calculated using *PCA_NetEmd* with size 4 graphlets and 90% explained variance. 221

E.6 COVID-19 forum - threads of each sentiment - alternative data granularity. 222

E.7 COVID-19 forum - posts in threads of each sentiment - alternative data granularity. 223

E.8 COVID-19 subforum - users in threads of each sentiment - alternative data granularity. 224

G.1 Comparison of testing strategies against reduction of contacts for different growth rates with POLYMOD contact patterns. The x-axis represents the number of total infections, averaged over 100 runs. The y-axis shows the maximum peak of new daily infections, averaged over 100 runs. The colour of each marker indicates the size of the regular tester cohort for the simulations with asymptomatic testing or the percentage of contacts removed from non-household layers for the simulations with contact reduction. 234

G.2 Comparison of testing strategies against reduction of contacts for different growth rates with CoMix contact patterns. The x-axis represents the number of total infections, averaged over 100 runs. The y-axis shows the maximum peak of new daily infections, averaged over 100 runs. The colour of each marker indicates the size of the regular tester cohort for the simulations with asymptomatic testing or the percentage of contacts removed from non-household layers for the simulations with contact reduction. 235

G.3 Impact of increasing the uptake on isolation on symptoms, daily symptomatic testing and PCR testing after being traced as a contact of a positive case. Numbers shown are the percentage decrease in the total number of infections, compared to 0% uptake on these interventions, for 0.05 and 0.1 growth rates in POLYMOD contact pattern. The top line on the x axis represents the uptake to testing after contact tracing, the bottom line represents the daily probability to testing when symptomatic. The y axis represents uptake to isolation on COVID or ILI symptoms. 236

G.4 Impact of increasing the uptake on isolation on symptoms, daily symptomatic testing and PCR testing after being traced as a contact of a positive case. Numbers shown are the percentage decrease in the total number of infections, compared to 0% uptake on these interventions, for 0.15 and 0.225 growth rates in POLYMOD contact pattern. The top line on the x axis represents the uptake to testing after contact tracing, the bottom line represents the daily probability to testing when symptomatic. The y axis represents uptake to isolation on COVID or ILI symptoms. 237

G.5 Comparison of LFD against PCR tests for symptomatic or contact traced individuals, for different values of asymptomatic testing uptake and different growth rates, with contact patterns obtained from POLYMOD. The y-axis shows the total number of infections. The x-axis represents the size of the regular asymptomatic tester cohort. Markers shown represent the median of 100 simulations and the shaded area represents the interquartile range. The policy used for the regular tester group is testing once weekly. 238

G.6 Comparing LFD and PCR tests and the impact of increasing the uptake of testing after being traced as a contact of a positive case and the daily probability of testing if symptomatic, for growth rates of 0.025, 0.05 and 0.1 with contact patterns obtained from POLYMOD. The y-axis shows the total number of infections. The bottom line on the x axis represents the uptake to testing after contact tracing, the top line represents the daily probability to testing when symptomatic. Color indicates the type of test and the adherence to isolation on symptoms. Markers shown represent the median of 100 simulations and the shaded area represents the interquartile range. 239

G.7 Comparing LFD and PCR tests and the impact of increasing the uptake of testing after being traced as a contact of a positive case and the daily probability of testing if symptomatic, for growth rates of 0.15, 0.225 and 0.3 with contact patterns obtained from POLYMOD. The y-axis shows the total number of infections. The bottom line on the x axis represents the uptake to testing after contact tracing, the top line represents the daily probability to testing when symptomatic. Color indicates the type of test and the adherence to isolation on symptoms. Markers shown represent the median of 100 simulations and the shaded area represents the interquartile range. 240

G.8 Comparison of number of tests required per person when using LFD or PCR tests for symptomatic or contact traced individuals, for 0.025, 0.05 and 0.15 growth rates, with contact patterns obtained from POLYMOD. The y-axis shows the mean number of tests per person, averaged over 100 runs. The bottom line on the x axis represents the uptake to testing after contact tracing, the top line represents the adherence to isolation on symptom onset. 241

G.9 Comparison of number of tests required per person when using LFD or PCR tests for symptomatic or contact traced individuals, for 0.15, 0.225 and 0.3 growth rates, with contact patterns obtained from POLYMOD. The y-axis shows the mean number of tests per person, averaged over 100 runs. The bottom line on the x axis represents the uptake to testing after contact tracing, the top line represents the adherence to isolation on symptom onset. 242

G.10 Impact on person-days of isolation of increasing adherence to the different testing strategies, for different growth rates with contact patterns from POLYMOD, when using PCR (Figure G.10a) or LFD (Figure G.10b) as the test for symptomatic or contact traced individuals. The x-axis represents the size of the regular tester cohort for the simulations with asymptomatic testing or the percentage of contacts removed from non-household layers for the simulations with contact reduction. The y-axis shows the mean number of person days of isolation per population members over the 180 days of simulation, averaged over 100 runs. These include days in isolation waiting for test results (amongst true positives and true negatives); isolation days amongst those testing positive who were true positive; and isolation days among those testing positive who were false positives. The colour of each marker indicates the percentage of days spent in isolation while infected. 243

G.11 Impact on person-days of isolation of increasing adherence to isolation on symptoms, daily probability of testing and adherence to testing after contact tracing, for 0.025, 0.05 and 0.1 growth rates with contact patterns from POLYMOD, comparing the outcome when testing with PCR or LFD. The bottom line on the x axis represents the uptake to testing after contact tracing, the top line represents the adherence to isolation on symptom onset. The y-axis shows the mean number of person days of isolation per population members over the 180 days of simulation, averaged over 100 runs. These include days in isolation waiting for test results (amongst true positives and true negatives); isolation days amongst those testing positive who were true positive; and isolation days among those testing positive who were false positives. The colour of each marker indicates the percentage of days spent in isolation while infected. 244

G.12 Impact on person-days of isolation of increasing adherence to isolation on symptoms, daily probability of testing and adherence to testing after contact tracing, for 0.15, 0.225 and 0.3 growth rates with contact patterns from POLYMOD, comparing the outcome when testing with PCR or LFD. The bottom line on the x axis represents the uptake to testing after contact tracing, the top line represents the adherence to isolation on symptom onset. The y-axis shows the mean number of person days of isolation per population members over the 180 days of simulation, averaged over 100 runs. These include days in isolation waiting for test results (amongst true positives and true negatives); isolation days amongst those testing positive who were true positive; and isolation days among those testing positive who were false positives. The colour of each marker indicates the percentage of days spent in isolation while infected. 245

Abstract

Complex networks are topologically rich graphs, containing properties that are neither regular nor purely random. Complex networks model real complex systems, so analyzing and extracting information from networks provides knowledge about the system itself. Network comparison is a task within network analysis aimed at quantifying similarities and differences between networks. Graphlets are small patterns of connections whose distribution in the network has been shown to be one of the most informative properties of the structural organization of the network, hence have recently been increasingly adopted as features in network comparison algorithms and measures.

NetEmd is a network comparison measure that uses graphlets as features for comparison, being able to identify random networks generated from the same random graph model or real networks from the same domain more accurately than previous network comparison measures for undirected networks, even when their size differs by multiple orders of magnitude. This thesis presents an extension of NetEmd to directed networks and a framework of noise reduction in graphlet frequencies prior to the using the NetEmd measure. Extensive testing is carried out on the performance of the extension and the noise reduction framework, showing the value of these techniques.

Network comparison based on graphlets has seen the majority of its use when applied to networks from the biology domain. From an application perspective, this thesis aims to demonstrate how insights based on network comparison with graphlets can be useful in other domains:

- Analysis of controversial discussion topics in online social networks often includes a component of sentiment labeling, detection or analysis. However, sentiment labeling is an expensive and labourious process, automatic sentiment detection shows poor performance in some domains and sentiment analysis cannot be performed without either of these. Considering vaccination discussions in online social networks in particular, this thesis shows how differences in network structure detected by network comparison are correlated with changes to

the sentiment towards vaccination. Network comparison can be used as a tool to guide when to perform sentiment analysis as well as a technique to enhance the conclusions from sentiment analysis.

- During the SARS-CoV-2 pandemic, social mixing prevention policies introduced changes in social contact networks in a way to prevent the spread of the virus. By generating networks based on age-mixing matrices obtained from social surveys representative of the UK population, this thesis performs an investigation of Test-Trace-Isolate strategies within an agent based model used to inform government policy, under different assumptions of social contact network structure.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

An undertaking as big as a PhD requires a strong support system and I am very fortunate to have been the recipient of much support over the past 5 years, so I would to take this time to thank everyone who helped me.

I want to begin by thanking my family who provided for all I have needed throughout my life. To my Mum and Dad, who continuously give me the love, nurture and even scientific advice that I needed to get to where I am. Words cannot express how lucky and privileged I feel to be your son. To my brother, who is always one step ahead of me and shows me a light to guide my path; if I am fortunate to achieve as much as you do, I know I will find happiness. To Néné, thank you for all the cute videos, the Sunday lunches and the literary inspiration to read more and be more knowledgeable, ultimately making writing this thesis an easier task. To Vóvó Fátima, who makes me want to be the best grandson in the world. To Paula and Emanuel, who are always eager to help me. To Rodrigo, I know you will grow up to make your cousin proud. To my grandparents who are no longer with us, I know you would be proud of me.

To my supervisors, Caroline and Thomas, who have taught me so much and given me a completely different perspective on what it means to do research and how to think about my own research. I consider myself to be extremely lucky to have been supervised by two people who are not only scientifically brilliant but also great human beings, who always made me feel safe and confident in my research after every meeting. I would also like to thank my collaborators over the past years, many of whom contributed with advice in this thesis, in particular I would like to thank Robert Gaunt, Luis Ospina-Forero, Rigina Skeva and Elizabeth Fearon. I would also like to thank Pedro Ribeiro for introducing me to the wonderful world of network science.

The third pillar that supported this endeavour were my friendships, as a source of entertainment, advice and inspiration. In Manchester, I am thankful for the many like minded people I befriended, with a special thanks to Mark, Ed, Luca, Jonathan, Lukas, Julia, Aitor and Aileen. In Porto, I left many friends that continue to be there for me.

Catarina, the light that brightens my days; Afonso, João Vitor, Ricardo, Salvador and Tiago, who were always the first ones to welcome me back; Marvin and Rafaela, who I always want to impress; and João Resende and Patricia, whose growth makes me proud and makes me strive to be better.

Lastly, I want to thank Adam, who put up with me almost every single day for 4 years, including a world pandemic with 3 lockdowns. Our philosophical discussions taught me a lot about myself, about life and who I want to be. Being flat mates with you has made taking a respite from work a much fuller experience and all the adventures, the joy and the sadness has shaped the way I am now and the way I think, ultimately influencing this thesis.

To all of you, my most heartfelt thanks.

Acronyms

ABM agent-based model 50, 55, 127, 147

ARI adjusted Rand index 11, 12, 65, 181–184, 192–194

AUPR area under precision-recall curve 10, 11, 65, 179–184, 189–191

DGCD Directed Graphlet Correlation Degree 8, 10–12, 58, 70–72, 180, 183, 184, 186, 188, 190, 191, 193, 194

EMD Earth Mover's Distance 43, 58, 75–77

FN False Negative 179

FP False Positive 179

GCD Graphlet Correlation Degree 7, 10, 11, 43, 44, 69, 189, 192

GCM Graphlet Correlation Matrix 44

GDA Graphlet Degree Distribution Agreement 7, 8, 10–12, 43, 45, 58, 60, 69, 71, 72, 180, 183, 184, 186, 189–194

GDD Graphlet Degree Distribution 45

ICA Independent Component Analysis 59, 64, 65, 70–72, 75–78, 149, 153, 180, 181, 183, 187–189, 191, 192, 194

ILI Influenza-like Illnesses 16, 131–133, 138, 139, 143, 145, 230–232

LFD Lateral Flow Device 16, 19–22, 126–128, 132, 133, 138, 140–147, 153, 230–232, 238–245

LFT Lateral Flow Test 37, 126, 127, 147, 148, 150, 153, 230

MMR measles mumps rubella 83, 88, 100, 199

NAbs Neutralising Antibodies 128, 147

ODE ordinary differential equation 50–52

PCA Principal Component Analysis 58, 59, 63, 65, 71, 75, 76, 78, 149, 153, 180, 183, 189, 191, 192, 194

PCR Polymerase Chain Reaction 16, 19–22, 37, 126, 128, 131–134, 138–144, 146, 147, 150, 153, 230–232, 236–245

RAT Rapid Antigen Test 126

RQ Research Question 36, 37, 82, 123, 124

SNAP Stanford Network Analysis Project 68

TN True Negative 179

TP True Positive 179

TTI Test-Trace-Isolate 36, 126, 127, 132, 150

UK United Kingdom 34–36, 80, 81, 101, 102, 125–127, 129–132, 144, 150, 152

Chapter 1

Introduction

Real world systems are sets of entities that, through their interactions or relationships, form a unified whole. These systems are often *complex*, in the sense that properties emerge which would otherwise be unpredictable from the behaviour of each individual part. Interactions between entities of a complex system can be modeled through graphs, giving rise to *complex networks*, graphs whose topological features are non-trivial, being neither regular nor purely random. Complex networks are found in the real world in a wide range of domains, such as brain networks [168], protein-protein interaction networks [75, 156], scientific collaboration networks [92, 115], social networks [88, 198] or infrastructure networks [121].

Developing methodology to analyze and extract information from complex networks is crucial to understanding the phenomena that originate the complex system they model. Thus, the study of complex networks has developed into an active area of scientific research [3, 114], particularly in the past 25 years as a result of advances in computing power and increased amount of data generation. The analysis of complex networks employs a variety of techniques, such as community detection [43] or node centrality [44]. A relatively new method for network analysis are network motifs [109], defined as "patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks" [109]. Network motifs make up a powerful tool for network analysis because they can be interpreted as units of functional behaviour encoded in the network [102]. Network motifs have been widely utilized in characterizing network from many domains [22, 175], but especially in biology [158, 168, 194]. However, by their definition, the frequency of network motifs is compared against a set of random networks generated by a random network

model, which constitutes a null model to compare against in order to determine the significance of a pattern, and different null models have been shown to produce different motifs for the same network [154, 161].

One open problem in network science where network motifs have enjoyed particular success is *network comparison*. The network comparison problem arises from the need to quantify differences and similarities between networks, particularly when their sizes differ but the networks are hypothesized to be related. Applications of network comparison include goodness of fit tests for random graph models of real world networks [108, 133] or in biological networks [4, 7]. Measures of distance between networks based on the distribution of *graphlets* [132] have risen to prominence recently [7, 152, 180, 192], as graphlets carry the same functional properties of network motifs, without the need to determine significance based on null models.

In this thesis, we advance network comparison methodology based on graphlets and apply it in two areas relevant to the COVID-19 pandemic. In particular, we focus on the applications of network comparison to the analysis of social networks, a domain in which network comparison based on network motifs and graphlets has seen little use, despite the advancements in biology networks.

Social networks are pervasive in modern society and thus have become fertile ground for the spread of misinformation about a multitude of topics, one of which is vaccination [13, 17]. The World Health Organization named ‘Vaccine hesitancy’, i.e. the reluctance or refusal to vaccinate despite the availability of vaccines, as one of the 10 major threats to global health in 2019 [187]. The issue of vaccine hesitancy has become particularly timely during the SARS-CoV-2 global pandemic, the virus that provokes the COVID-19 disease, where the vaccine has been instrumental in preventing deaths and hospitalizations due to the virus, in addition to ending “lockdown" style policy measures that carry impacts beyond global health. In spite of this success, there is still a high level of vaccine hesitancy within specific communities, such as minority ethnic groups [146] and those from most deprived areas [24].

The majority of research directed towards vaccination discussions in online social networks tends to focus on the most popular media, such as Facebook [66, 155] or Twitter [1, 11, 98, 134]. However, it has been shown that *online discussion boards* are used as sources of information about immunisation and parents who use these discussions forums were more likely to question the decision to vaccinate than those who use Facebook or Twitter [20], yet not as much research as been dedicated to studying them. One popular online chat board is contained within the United Kingdom (UK)

parenting website Mumsnet [145]. Mumsnet generates more discussion content and with a greater variety of opinions regarding vaccination than other discussion boards more focused exclusively on vaccination [165].

The most common way of analyzing vaccination discussions in online social networks is through sentiment analysis [112, 150, 166], which requires annotating discussions according to their view on vaccination. This process is extremely laborious when performed through manual annotation and automatic detection of sentiment does not achieve adequate performance on Mumsnet data [166]. By representing user interactions over time in Mumsnet through series of complex networks, we show how using network comparison to detect differences in the structure of interactions can be a sign of change in sentiment towards vaccination. As the differences in network structure are identified as a result of the characteristics of graphlets, the units of functional behaviour, combining sentiment analysis with network comparison allows the generation of hypothesis regarding changes to user behaviour when discussing vaccinations.

The SARS-CoV-2 pandemic, the virus responsible for the COVID-19 disease, has been an extremely impactful event worldwide throughout 2020, 2021 and 2022. In the UK, Government's response to the pandemic has been advised by the Scientific Advisory Group for Emergencies (SAGE) and sub-committees, such as the Scientific Pandemic Influenza Group on Modelling (SPI-M). As part of Test, Trace and Isolate COVID-19 rapid response modelling project which provided advice to SPI-M, another application of network comparison we explore in this thesis is studying changes in the structure of social contact networks due to the COVID-19 pandemic. In the initial phase of the COVID-19 pandemic, infection control to mitigate the impacts of the disease was achieved by policy aimed at the reducing the amount of contact between people. Social surveys aimed at understanding the impact of these measures on the amount of daily contacts in the population show that, by 2022, the average number of contacts still had not reached pre-pandemic levels [77, 78]. Mathematical modelling of COVID-19 transmission and control measures has been an instrumental tool in advising governments about potential outcomes of policy changes [45, 120, 171]. Network based models, those where infection spread is structured according to a network, need to incorporate this change in contact patterns observed in the population, in order to provide realistic outcomes of infection control measures.

One such model is *Covasim* [82], an agent based model with multiple layers of social contact networks meant to mimic settings where transmission of COVID-19

can occur, such as work places, schools or households. It has been used at different points of the pandemic in the UK to evaluate the potential impact of policy proposed by the government [122–125]. We extend the models for network generation in Covasim by discriminating the number of contacts between people of different age groups, through age-mixing matrices obtained from social contact surveys, namely POLYMOD [111] and CoMix [78]. Using network comparison, we show how the patterns of social contact change when generating networks based on age-mixing matrices, which affect chains of transmission for the virus. We present an experimental analysis of Test-Trace-Isolate (TTI) interventions using these networks generated from age-mixing matrices, an analysis that was used as advice to the UK government on the role of asymptomatic testing to mitigate transmission of COVID-19 and the impact of the disease.

1.1 Thesis contributions

The aim of this thesis is to develop new techniques for comparing complex networks using distributions of graphlets and show the applicability of this approach to new domains. To this end, we propose a set of research questions and delineate the contributions in this thesis to answer them:

Research Question (RQ) 1: Given a dataset of complex networks, how can their differences and similarities be quantified?

The contributions in this thesis towards RQ1 are two fold. In the case where the dataset is composed of directed networks, we extend NetEmd [180], a network comparison measure based on graphlet distributions in undirected networks, to support the comparing directed networks. We also propose a pre-processing step to the comparison where dimension reduction techniques are used as a denoising mechanism of the network comparison features, showing that this pre-processing step helps distinguishing both directed and undirected networks.

RQ2: What is the impact of COVID-19 on the structure of vaccination discussions in online discussion boards?

We analyze the characteristics of online discussion forum Mumsnet, comparing discussions before and after the COVID-19 pandemic. Through network comparison, we demonstrate that user interactions in the two periods of time are characterized by distinct patterns. We also find that significant deviations in network structure between networks generated from temporally contiguous data occur in the same periods of time

when there are changes in the sentiment towards vaccination.

RQ3: How do changes in the structure of social contact networks affect mathematical modeling of COVID-19 transmission?

We implement social contact networks generated from age-mixing matrices derived from social surveys POLYMOD [111] and CoMix [78] in a previously published agent based model for COVID-19 simulation, Covasim [82]. We observe that networks generated based on social contact during the pandemic are organized in a different manner compared to those based on social contact before the pandemic. Using these networks as the structure of contacts for COVID-19 transmission, we investigate different strategies for regular asymptomatic testing, the efficacy of interventions such as contact tracing and isolation in a scenario with no asymptomatic testing in the population and the impact of swapping Polymerase Chain Reaction (PCR) testing for a single Lateral Flow Test (LFT).

1.2 Thesis structure

The remainder of the thesis is divided into five chapters as follows:

- **Chapter 2:** an introduction to concepts and definitions related to network science used throughout this dissertation.
- **Chapter 3:** presentation of an extension to the network comparison measure NetEmd [180] to directed networks and a framework that uses dimension reduction methods as a way to denoise the features of NetEmd.
- **Chapter 4:** analysis of vaccination discussions in the Mumsnet forum. Using network comparison, the way users interact in these discussions threads are shown to vary over time and the differences to be correlated with changes to the sentiment towards vaccination.
- **Chapter 5:** evaluation of different asymptomatic testing strategies and the role of adherence to symptomatic testing, contact tracing and isolation on symptoms during an infection wave of COVID-19. This is done in an agent based model with social contact networks, whose connections are informed by age-mixing matrices in different transmission settings.
- **Chapter 6:** summary of findings from each chapter and potential directions for future research.

Chapter 2

Background

2.1 Network Terminology

This chapter serves as an introduction to network science related concepts and definitions used throughout the remainder of the thesis.

Entities of a complex system are represented in a complex network by objects called **nodes (or vertices)** and their interactions are represented through **edges (or connections)**. Examples of nodes and edges that we use in this work are users of an online social network that are connected if they participate in the same discussion; or people in a social contact network, whose connections signify the possibility of COVID-19 transmission.

Graphs are mathematical objects used to represent a complex network. A graph $G = (V(G), E(G))$ is composed of a set of nodes $V(G)$ and a set of edges $E(G) \subseteq V(G) \times V(G)$, where each edge is represented by a pair $(u, v) \in E(G)$ for $u, v \in V(G)$. A graph can be either **directed**, when the order of the vertices in the pairs expresses direction, meaning that $(u, v) \in E(G)$ does not imply that $(v, u) \in E(G)$, or **undirected**, otherwise. Graphs are assumed to be labeled, all nodes are assigned consecutive integer numbers starting from 0 and running to $|V(G)| - 1$. The **size** of a graph is the number of vertices in the graph, written as $|V(G)|$. A size-k graph is a graph with k vertices. The **density** of a graph is the portion of edges present in the graph over the total potential ones. Calculated as $|E(G)|/\binom{|V(G)|}{2}$ in undirected networks and $|E(G)|/(|V(G)|^2 - |V(G)|)$ in directed networks. In directed networks, a **reciprocal edge** is one such that $(u, v) \in E(G) \wedge (v, u) \in E(G)$. The **reciprocity** of a network is the proportion of edges that are reciprocal.

Graphs can be classified into different types according to constraints or additions

to the set of nodes and edges. A graph is called **simple** if it does not contain multiple edges (two or more edges connecting the same pair of vertices) or self-loops (an edge of the form (u,u) that connects a vertex to itself). In this work, we consider only simple graphs. A graph is called **connected** if all nodes are reachable from each other. A node $v \in V(G)$ is reachable from a node $u \in V(G)$ if $(u,v) \in E(G)$ or if there is a path between u and v, i.e., a sequence of edges of the form: $(u,t_1),(t_1,t_2),\ldots(t_n,v)$. A **bipartite graph** is a graph whose vertices represent two disjoint groups of entities and there are no edges between members of the same group. Formally, $G = (V(G_1) \cup V(G_2), E(G))$, such that $V(G_1) \cap V(G_2) = \emptyset$ and $E(G) \subseteq V(G_1) \times V(G_2)$, with each edge represented by a pair $(u,v) \in E(G)$ for $u \in V(G_1) \implies v \in V(G_2)$ or $u \in V(G_2) \implies v \in V(G_1)$. A **weighted graph** is a graph whose edges represent additional information, a **weight**, about the relationship between vertices, beyond the existence of that relationship. The most common example of an edge weight is a numeral attribute representing, for instance, the frequency of the relationship. In the case of a numerical attribute, the set of edges is defined as $E(G) \subseteq V(G) \times V(G) \times \mathbb{R}$ and each edge is represented as a triple $(u,v,w) \in E(G)$ for $u,v \in V(G)$ and $w \in \mathbb{R}$.

The **neighbourhood** of a vertex $u \in V(G)$ is the set of nodes that u is connected to and defined as $N(u) = \{v : (v,u) \in E(G) \vee (u,v) \in E(G)\}$. The **degree** of a vertex is the number of edges it participates in, which is equivalent to the size of the node's neighbourhood, $|N(u)|$. A node is called **isolated** if it has no connections, i.e., its degree is 0. In directed networks, the degree of a vertex can be split into the **in-degree** and **out-degree**, counting the number of incoming and outgoing edges, respectively, from that node. In directed networks, the degree of a vertex is the sum of its in-degree and out-degree. In undirected networks, there is no distinction between between in-degree and out-degree, so the degree of a node u is simply $|N(u)|$. The **degree sequence** the sequence of node degrees in a network. A sequence s is called a **graphical sequence** if exists a graph whose degree sequence equals s. The **weighted degree** of a node is the sum of weights of all edges the node participates in. In directed networks, the **weighted in-degree** and **weighted out-degree** of a node are the sum of weights of all incoming and outgoing edges, respectively, from that node.

A **subgraph** G_k of a graph G is a size-k graph such that $V(G_k) \subseteq V(G)$ and $E(G) \supseteq E(G_k) \subseteq V(G_k) \times V(G_k)$ and is called **induced** if $\forall u,v \in V(G_k),\ (u,v) \in E(G_k) \leftrightarrow (u,v) \in E(G)$. Two graphs G and H are **isomorphic**, written as $G \sim H$, if there is a bijection between $V(G)$ and $V(H)$ such that two vertices are adjacent in G if and only if their correspondent vertices in H are adjacent. A **match** of a graph H in a larger

graph G is a set of nodes that induce the respective subgraph H. In other words, it is a subgraph G_k of G that is isomorphic to H. The frequency of a subgraph G_k is then the number of different matches of G_k in G.

Orbits are unique positions of a graph, calculated by partitioning the set of vertices into equivalence classes where two vertices belong to the same class if there is an automorphism that maps one into the other [139]. Orbits represent the different roles within a subgraph and encapsulate the idea that multiple nodes can perform the same function within a subgraph. For instance, in undirected subgraph G_1 of Figure 2.1, there are two orbits, one belonging to the centre node and another orbit shared by the extremity nodes. These two orbits have different roles within the subgraph in the sense that the centre node is responsible for connecting the extremities, who in turn share the role of supporting the centre node. This contrasts with undirected subgraph G_2 where are all nodes have the same orbit, carrying the role of representing a transitive relationship between the three nodes.

Graphlets are small, connected, non-isomorphic and induced subgraphs [132]. The smallest graphlet considered is a single edge, which can be seen as a size-2 subgraph. An undirected edge has a single orbit (three in the directed case) and the frequency of a node in this orbit is equivalent to the degree of the node. Figure 2.1 shows the graphlets of size 2, 3 and 4 in undirected networks and size 2 and 3 in directed networks, alongside the respective orbits. The **graphlet degree vector** is an extension of the definition of degree to size-k graphs, representing how many times a node occurs in each orbit. The **graphlet degree matrix** is the collection of graphlet degree vectors of each node in a graph.

A graph where all nodes are connected to each other is called a **clique**, or a **complete graph**. In other words, it is a graph with density of 1. In Figure 2.1, corresponds to undirected graphs G_2 and G_8 and directed graph G_{14}. The **clustering coefficient** of a graph is a measure of the amount of transitivity in the graph. It is calculated by dividing the number of triangles (cliques of size 3, corresponding to the undirected subgraph G_2 in Figure 2.1) by the total number of triples a node participates in.

A **connected component** of a graph is a connected induced subgraph that is not a part of another larger connected induced subgraph. In other words, it is a subgraph where all nodes are reachable from each other and there are no edges to any node outside this subgraph, therefore inducing a disjoint partition of the set of vertices. The connected component with the highest number of nodes is called the **largest connected component** or the giant component.

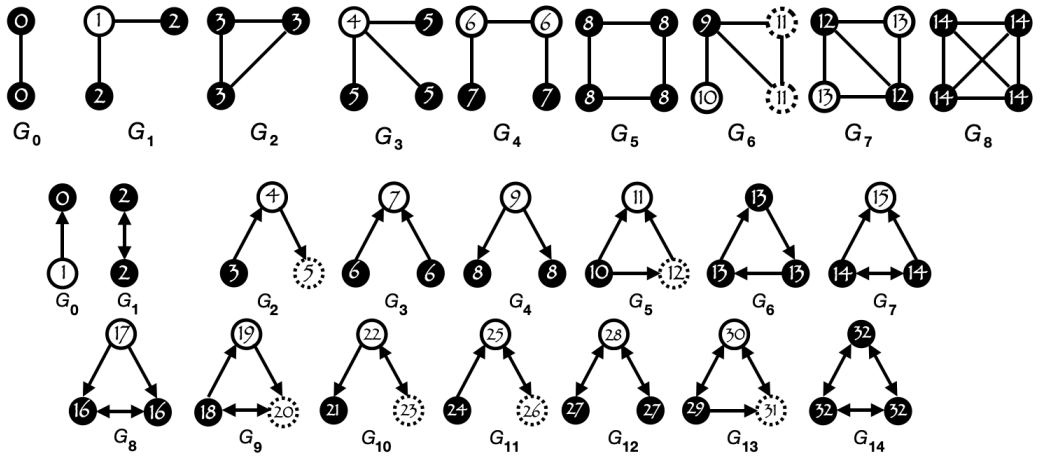


Figure 2.1: Undirected graphlets of size 2, 3 and 4 and directed graphlets of size 2 and 3. Nodes are numbered according to their orbit, counted by type (directed or undirected) from the size 2 graphlets, and nodes with the same number represent the same orbit. The background colour and shape of node boundary further emphasizes nodes belonging to the same orbit in the same graphlet.

Random graphs are graphs generated by a stochastic process. Different **random graph models** produce graphs with different properties meant to mimic properties of complex networks. A common model for generating random graphs is a prescribed degree distribution is the **configuration model**. Taking a graphical sequence as input, the model works by assigning each node a degree represented as a half-edge or a stub. Pairs of half-edges as chosen uniformly at random and connected to form an edge, iteratively until no stubs remain.

2.2 Network Comparison

Comparing different objects is a fundamental part of human cognition, therefore it is only natural that analysis of complex networks encompasses a comparison element. It is trivial to know if two networks are different, one needs only to look at the set of nodes and edges that compose the networks and claim a difference if these sets do not match. On the other hand, knowing if two networks are the same up to a relabeling function on the node set is known as the graph isomorphism problem, which has been shown to be in the NP complexity class [26]. Between these two extremes, **network comparison** has been established as an area of study within network analysis that combines network statistics and computable properties of a network to tell how similar or dissimilar two networks are. Formally, we define network comparison as:

Problem definition - Network Comparison. *Given a set of networks* $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$ *and a set of features* $f_i \in \mathcal{F}$*, compute the distance* $d_{f_i}(G,H) = d(p_{f_i}(G), p_{f_i}(H))$ *for every pair of networks* $G,H \in \mathcal{G}$ *and for every feature* $f_i \in \mathcal{F}$*, where* $p_{f_i}(G)$ *is the distribution of feature* f_i *in G.*

Network comparison has been the subject of an increasing amount of research, see for example [4, 5, 7, 8, 22, 84, 100, 108, 118, 132, 152, 169, 180, 190, 192]. A category of methods for network comparison focuses on solving a sub-problem called **network alignment**, where the goal is to create a mapping between the nodes of the networks being compared [6, 41, 57, 86, 101, 107, 164].

In this work we focus only on **alignment-free** methods, that use global features of the networks to determine similarity or dissimilarity. Among these, methods that use graphlets have emerged as the state of the art in the area, in particular, we highlight NetEmd [180], TriadEMD [190], Graphlet Degree Distribution Agreement (GDA) [7, 132] and Graphlet Correlation Degree (GCD) [152, 192].

2.2.1 NetEmd

NetEmd [180] is a network comparison measure that relies on structural features of the network, mainly the distribution of orbit frequencies. The core idea is formalizing the intuition that the shape of the degree distribution is indicative of the network's generation mechanisms, for instance, a network with a power law degree distribution is generated by a process distinct from a network with a uniform degree distribution. As the graphlet degree vector is a generalization of the degree distribution for graphlets of size $k \geq 3$, the shapes of the distributions of each orbit also carry information about the topology of the network. Note that because the shape of a distribution is invariant under linear transformations such as translations, using the shape as the focus of the comparison is well-suited to comparing networks of different sizes and densities.

Wegner et al. [180] postulate that "any metric that aims to capture the similarity of shapes should be invariant under linear transformations of its inputs." Thus, they define a measure of similarity between distributions p and q, with non-zero and finite variances, using the Earth Mover's Distance (EMD) [149]:

$$EMD^*(p,q) = \inf_{c \in \mathbb{R}}(EMD(\tilde{p}(\cdot + c), \tilde{q}(\cdot))),$$

where $\tilde{p}$ and $\tilde{q}$ are the distributions resulting of scaling p and q to variance 1. Any distance metric d can be used to generate d^*; EMD was used by the authors as it has

been shown to be an appropriate metric to compare shapes of distribution in domains such as information retrieval and it produced better results than other distance metrics, like the Kolmogorov or L^1 distances.

Given two networks G and H and a set of m orbits $O = \{o_1, o_2, \ldots, o_m\}$, the *NetEmd* measure is defined as:

$$NetEmd_O(G,H) = \frac{1}{m}\sum_{i=1}^{m} EMD^*(p_{o_i}(G), p_{o_i}(H)), \tag{2.1}$$

where $p_{o_i}(G)$ and $p_{o_i}(H)$ are the distributions of orbit i in graphs G and H, respectively. Note that O may be replaced by any set of network features; in this work we focus on orbits only.

Xu and Reinert [190] present an extension of NetEmd to directed networks, keeping Equation 2.1 the same and using orbits 3 to 32 from Figure 2.1 as the set of orbits O.

2.2.2 GCD

Yaveroğlu et al. [192] note that orbit counts have dependencies between them, making them redundant since they can be expressed as a linear combination of other orbits. The authors identify 11 out of 15 orbits in graphlets of size up to 4 and 56 out of 73 orbits in graphlets of size up to 5 as non-redundant. The authors construct the graphlet degree vector of each node in the network using this reduced number of orbits (although the full set of orbits may also be used), making up a matrix where each row is the graphlet degree vector of each node, and compute the Spearman's correlation coefficient between each pair of orbits, i.e., each pair of columns in this matrix. The pairwise Spearman's correlation coefficients are aggregated in a square $m \times m$ matrix called the *Graphlet Correlation Matrix (GCM)*, where m is the number of orbits used (e.g., 11 if using non-redundant orbits of graphlets up to size 4).

To compare two networks, the authors propose the *GCD*, which is the Euclidean distance between the upper triangle of the GCM of each network.

Sarajlić et al. [152] extend GCD to directed networks, finding 13 non-redundant orbits in graphlets of size up to 3 and 129 in graphlets of size up to 4.

2.2.3 GDA

Przulj et al. [132] introduced the Graphlet Degree Distribution (GDD), calculated as follows. Let $d_G^o(k)$ be the number of nodes of graph G that participate k times in orbit o, i.e., $d_G^o(k)$ is graphlet degree distribution of orbit o. The authors scale the distribution to decrease the contributions of larger orbits by calculating $S_G^o(k) = d_G^o(k)/k$ and then normalize the distribution as $N_G^o(k) = S_G^o(k)/\sum_{k=1}^{\infty} S_G^o(k)$. To compare the GDD distributions of the same orbit in different networks G and H, the authors propose the GDD-agreement (GDA) metric defined as:

$$A^o(G,H) = 1 - \left(\sum_{k=1}^{\infty} [N_G^o(k) - N_H^o(k)]^2 \right)^{\frac{1}{2}}.$$

As a measure of *agreement*, the output of GDA is 1 if the distributions are identical, as opposed to NetEmd which is a measure of *distance* meaning it will output 0 for identical distributions. To aggregate the agreements of the multiple orbits, the authors propose either an arithmetic or a geometric mean. We compare against the implementation of Aparício et al. [7], which extended this idea to directed networks, using the arithmetic mean as the aggregation function and is available at [138]. This implementation uses the idea that if a graphlet has a frequency of 0 in both networks, then the orbits from that graphlet are excluded from the arithmetic mean, for both in the directed and undirected versions.

2.2.4 Applications of network comparison

Network comparison as a technique to analyze complex networks has been applied in a wide variety of domains. The majority of its applications can be found in biology, metabolic networks [157, 200], protein-protein interaction networks [59, 87] or gene regulatory networks [15, 104] are examples of areas within biology that have benefited from insights obtained by using network comparison.

But even outside of biology, network comparison has been widely applied. In one of the seminal network motifs papers, Milo et al. [108] show how comparing networks by their network motif signature leads to groups of networks that share a common domain. In economics, Sarajlić et al. [152] analyze the evolution of world trade networks over time. In the social domain, Choobdar et al. [22] compare co-authorship networks from different scientific fields to discover similar patterns of publication in scientific fields that were not expected to be alike. Also in the social domain, McMillan and

Felmlee [106] compare different types of social networks to uncover patterns that are common to each type of social network.

2.3 Counting graphlets

In order to compare networks based on the distribution of graphlets, one needs to compute the graphlet degree vectors, which involves determining the frequency of each graphlet. This computational task is related to the subgraph isomorphism problem: given two graphs G and H, determine if G contains a subgraph isomorphic to H. The subgraph isomorphism problem is a known NP-complete, the proof that this problem belongs to this complexity class was presented by Cook in his seminal paper [26].

Counting subgraphs has become a particularly active area of network science after the introduction of network motifs in 2002 [109]. Ribeiro et al. [139] has recently published a systematic review of algorithms proposed to efficiently determine the frequency of subgraphs. The authors discriminate between algorithms that compute the frequency solely of subgraph or “orbit-aware" methods that are able to orbit frequencies for each node as well. We now give a brief overview of ORCA [65], the algorithm employed by NetEmd [180] to obtain orbit counts that only works in undirected networks, and G-Tries [7, 140], a data structure that supports an algorithm of the same name, which,to the best of our knowledge, shows the best performance in directed networks.

2.3.1 ORCA

ORCA [65] is an analytical approach to counting undirected graphlets, in the sense that it avoids enumerating explicitly sets of k nodes and performing an isomorphism test to determine which subgraph the set is isomorphic to. This category of approaches has been shown to compute subgraph frequencies orders of magnitude faster than enumeration approaches, as well as scaling up to larger graphs. ORCA is also considered to be a matrix based method [139], as it builds a matrix of linear equations that relate the frequencies of each subgraph with other subgraphs of smaller or equal size. In practice, ORCA achieves this by counting orbits instead of subgraphs directly, constructing the system of linear equation to have rank equal to the number of orbits minus one. Then, by efficiently enumerating a single orbit (in the case of 4 nodes graphlets, orbit 14 in Figure 2.1), the system can be solved from this single orbit. An important

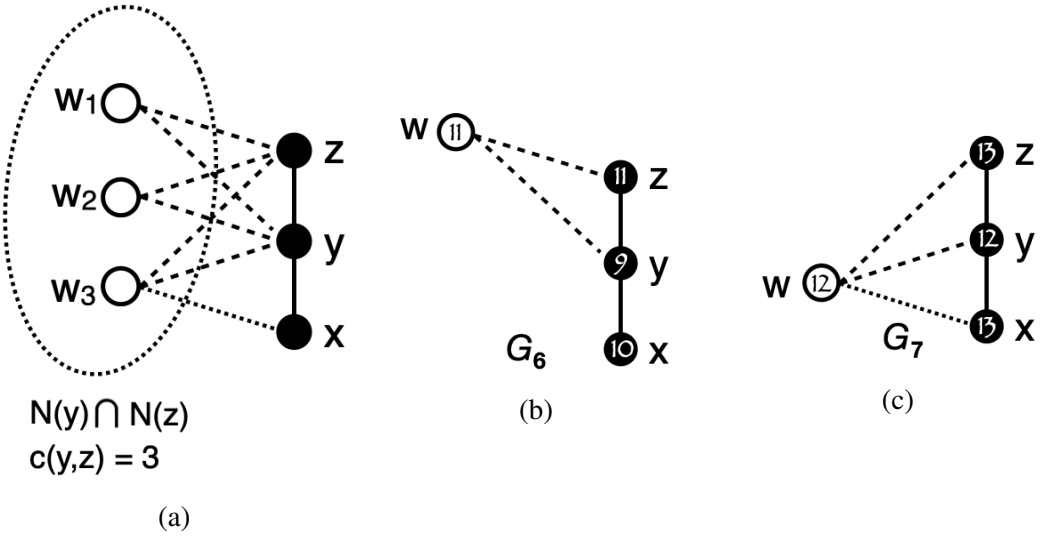


Figure 2.2: Figure adapted from Hočevar and Demšar [65], showing the relationship between undirected orbit 10 and 13 from Figure 2.1. In Figure 2.2a, we see the graphlet being extended, containing nodes x, y and z, whose edges are marked by a solid line. Each of w_i is a candidate node for the extension as they are the nodes in the neighbourhood of both y and z, their edges to these nodes exist by definition and are represented through a dashed line. Finally, the dotted line is optional. Its absence leads to a four node graphlet isomorphic to G_6, as pictured in Figure 2.2b. On the other hand, its existence leads to a four node graphlet isomorphic to G_7, as pictured in Figure 2.2c.

term in the equations of this system, that constitutes the right hand side of the equations, are terms computed from properties of the graph, like node degree. In particular, for the examples we show in this section, let $c(u,v) = |N(u) \cap N(v)|$ denote the size of the intersection between the neighbourhoods of two nodes, u and v, that is number of common neighbours of the two nodes

The main idea behind ORCA when counting graphlets with 4 nodes (note that this idea also holds for larger subgraphs) is that if a node participates in some graphlet G_i with 4 nodes, then it also necessarily participates in a graphlet G_j with 3 nodes. By observing that all 4 node graphlets can be constructed by adding a node to a 3 node graphlet, the authors show that to determine the orbit frequencies in 4 node graphlets of a node, it is only necessary to enumerate size 3 graphlets and determine their possible extensions.

To illustrate this idea, we show the example that Hočevar and Demšar [65] demonstrate in their original work, relating the frequency of orbit 10 and 13 from Figure 2.1, in Figure 2.2.

In this example, we consider the existence of a graphlet induced on nodes x, y and

z which is isomorphic to G_1. In this example, a proposed extension of this graphlet is given via a node w connected to both y and z, of which there are $c(y,z)$ candidates (equal to 3 in Figure 2.2a). This extension leads to two possible four node graphlets, depending on the existence of the edge (w,x). If the edge is present, then the graphlet is isomorphic to G_7 and x may be found in orbit 13; on the contrary, the resulting graphlet is isomorphic to G_6 and x belongs to orbit 10. As all the nodes in $N(y) \cap N(z)$ must participate in G_6 or G_7, establishing x in orbit 10 or 13, one can write $o_{10} + o_{13} = c(y,z)$ for this particular triplet x, y, z. This sum is then done over all 3 node paths (G_1) starting at node x, while taking symmetries into consideration as both G_6 and G_7 each appear twice with z and w and y and w with roles reversed, respectively. The final equation becomes:

$$2o_{10} + 2o_{13} = \sum_{\substack{y,z:x,z \in N(y) \\ G[x,y,z] \sim G_1}} c(y,z),$$

where $G[x,y,z]$ is the subgraph induced on x, y, z. The remaining 9 equations can be found in the supplementary material of the original work by Hočevar and Demšar [65].

2.3.2 G-Tries

A G-Trie [140] is a data-structure to designed to efficiently count subgraphs. It is a prefix tree (*trie*) for graphs, where the root holds a single node and each successor of a node in the trie is obtained by adding a vertex to the graph represented in its parent node. Unlike ORCA, the subgraph counting algorithm based on G-Tries (dubbed *GT-Scanner)* is an enumeration approach, meaning that all subgraph occurrences are generated and are categorized according to their isomorphism class. The G-Trie data structure allows an efficient combination of these two steps to prevent repeated isomorphism tests, which achieved a speed up of up to two orders of magnitude over previous approaches. Recently, Aparício et al. [7] extended G-Tries to encompass information about orbits, demonstrating how to use this graphlet prefix tree to generate the graphlet degree vector of a network. GT-Scanner can be seen as a depth-first search algorithm that maps a graph onto the G-Trie.

Figure 2.3 shows an example of, given a small directed network, how GT-Scanner is used to determine the frequency of five directed graphlets, two of size 3 and 3 of size 4. We omit the creating process of the G-Trie, whose details can be found in [7], assuming that graphlets D, E and G were inserted in the G-Trie in this order, creating

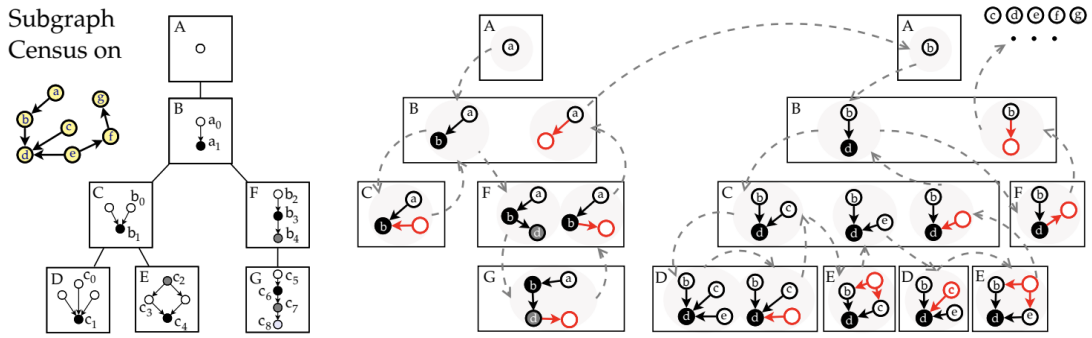


Figure 2.3: Figure taken from Aparício et al. [7]. Shows the subgraph census (counting the frequency of each subgraph) using a specific G-Trie. Nodes in red mean that no candidate was found for that particular graphlet and the dotted arrows represent the search path through the G-Trie.

intermediate tree nodes A, B, C and F.

GT-Scanner works by initially assigning all vertices to the root of the G-Trie, marked as A in Figure 2.3. Then, for each node, it traverses the G-Trie by generating candidates based on the connections of the current set of nodes. Taking the example of Figure 2.3, it starts by vertex a, whose only neighbour is vertex b, which matches the graphlet on G-Trie node B, increasing the frequency of orbit a_0 of vertex a and orbit a_1 of vertex b. In the next step of the algorithm, there is no valid vertex in the graph that fits G-Trie node C, so the search tracks back and matches G-Trie node F instead, by expanding the current set of vertices with vertex d. As this matches a graphlet in the G-Trie, the frequency of orbits b_2, b_3 and b_4 are increased for vertices a, b and d, respectively. The traversal of the G-Trie continues, but there is no candidate vertex in the graph that matches G-Trie node G, so the search tracks back. Similarly, in G-Trie nodes F and B there are no alternative vertices in the graph to continue the search, so the process is restarted with vertex b in the root of the G-Trie.

2.4 Mathematical modelling of infectious diseases

We now provide a brief overview of modelling methods for infectious diseases, as background for Chapter 5. The models we focus on describing the dynamics of disease transmission between hosts, ignoring or underplaying the importance of within-host viral dynamics. The majority of this section was written using the book by Kiss et al. [83].

Mathematical models for infectious diseases come in a spectrum that represents

a trade-off between complexity and explainability for causality. Simple models, like classical compartmental models, allow encoding the main mechanisms of transmission that answer questions about the probable existence of an epidemic, as well as its size and length, but this simplicity is often acquired at the cost of accuracy. On the other end of the spectrum, complex agent-based models (ABMs) carry a greater predictive power by including individual-level interactions, but it is harder to create a link between the coded complex dynamics and the results observed. This section is organized through this increase in complexity, we start by describing simple compartmental models, how to increase their complexity by removing the assumption of a well-mixed population and finalize with ABMs.

2.4.1 Compartmental based models

This category of models works by splitting the population in labeled compartments that represent the status of each individual towards the disease in study. These models are formulated as Markov chains, as the transition probability between compartments depends only on the current compartment.

Deterministic compartmental models are described through ordinary differential equations (ODEs), whose variables represent the size of each compartment and their derivatives over time encode transmission properties of the disease.

2.4.1.1 Susceptible - Infected

We start by one of the most simple models, the *SIS* model. In this case, only two compartments are represented, dividing the population into those who are *susceptible* (S) or *infected* (I). Upon the end of an infection event, an individual returns to the susceptible compartment. This model is used for diseases that do not provide long term immunity after infection, like influenza, so individuals may become infected again.

The ODE system for this model is written as:

$$\frac{dS(t)}{dt} = \gamma I(t) - \beta I(t)\frac{S(t)}{N},$$
$$\frac{dI(t)}{dt} = \beta I(t)\frac{S(t)}{N} - \gamma I(t),$$

where $S(t)$ and $I(t)$ denote the size of the susceptible and infected compartments at time t, β and γ are positive constants that represent the infection and recovery rates and N is the size of the population.

The meaning of these equations is simple: the rate of change of individuals in susceptible status depends on the rate of recovery and the number of infected individuals (γI) minus the total rate of infectious contacts (βI) but only a fraction of which are susceptible individuals that can lead to a new infection ($\beta IS/N$); the rate of change of individuals in the infected compartment is the opposite.

In systems of equations such as the one presented above, it is of interest to consider the steady-state (or equilibria) of the system, that is the behaviour of the system stabilizes. In this context, the steady states of interest are the disease free state $I = 0$ or when the disease becomes endemic ($I = N(1-\gamma/\beta)$), meaning that the disease will keep propagating in the population without imported infections, which exists only if $\beta/\gamma > 1$.

2.4.1.2 Susceptible - Infected - Removed

Another simple compartmental model, perhaps the most classic example thereof, is the *SIR* model, which adds a new compartment to the above, representing the *removed* individuals. These are individuals that were infected and recovered from the disease or died. In the case where the model does not contain a transition from the removed compartment to susceptible, it can be considered a case where the disease provides long term immunity after infection. An alternative to this model is the *SIRS* model, where individuals may become susceptible to infection after recovery.

For the SIR model, the ODE system is written as:

$$\begin{aligned}
\frac{dS(t)}{dt} &= -\beta I(t)\frac{S(t)}{N}, \\
\frac{dI(t)}{dt} &= \beta I(t)\frac{S(t)}{N} - \gamma I(t), \\
\frac{dR(t)}{dt} &= \gamma I(t),
\end{aligned}$$

thus, the similarities with the SIS model are evident.

2.4.1.3 Susceptible - Exposed - Infected - Recovered

We end this short overview of compartmental models with the *SEIR* model, which inserts an intermediate status between susceptible and infected that represents a latency period until an individual becomes infectious (the E stands for *exposed*). We highlight

this model because Covasim's model [82], albeit not a compartmental model, implements a similar disease progression scheme, where individuals go through an exposure period before becoming infectious.

In the case of the SEIR model, the ODE system is written as:

$$\begin{aligned}
\frac{dS(t)}{dt} &= -\beta I(t)\frac{S(t)}{N}, \\
\frac{dI(t)}{dt} &= \beta I(t)\frac{S(t)}{N} - \alpha E(t), \\
\frac{dI(t)}{dt} &= \alpha E(t) - \gamma I(t), \\
\frac{dR(t)}{dt} &= \gamma I(t),
\end{aligned}$$

where α^{-1} is the average latency period, assuming that the latency period is a random variable with exponential distribution of parameter α.

2.4.1.4 Stochastic models

The deterministic models we presented in the previous sections rely on the assumption that the size of each compartment is big enough that mixing in the population is homogeneous, i.e., members of each compartment interact equally with members of all compartments. At the beginning of a disease outbreak, it is unlikely that this condition is observed as the majority of the population starts as susceptible and the amount of infected individuals is small. As we model under the assumption that all individuals are capable of infecting one another (the underlying assumption is that the contact network is complete, we show how including network structure changes our model in the following section), a more realistic model needs to incorporate stochastic effects.

Taking the SIS model as an example, the stochastic formulation is achieved by considering $S(t)$ and $I(t)$ as discrete random variables (note that because $S(t) = N - I(t)$, there is only one independent random variable). Infection is transmitted between a pair of susceptible and infected individuals at a rate τ, such that τN equals β, and each individual recovers at a rate γ independently from all other individuals. One possible way to model this process is treating it as a continuous-time Markov chain with transitions:

$$\begin{aligned}
(S,I) &\xrightarrow{\tau SI} (S-1, I+1), \\
(S,I) &\xrightarrow{\gamma I} (S+1, I-1).
\end{aligned}$$

This process is usually described in terms of Kolmogorov equations, to obtain the probability that j infected individuals are observed at time t as:

$$\frac{dp_j(t)}{dt} = \tau(j-1)(N-j+1)p_{j-1} - (\tau j(N-j)+\gamma j)p_j + \gamma(j+1)p_{j+1}.$$

Focusing on the initial stage of the epidemic, where we can assume that the number of susceptible individuals can be approximated by N, the epidemic will become extinct if the infection rate is smaller or equal to the recovery rate (the basic reproductive rate, R_0, is given by $\tau N/\gamma \leq 1$). On the other hand, if $R_0 > 1$ and there are i_0 initially infected individuals, the probability of an epidemic is given by $\mathbb{P}(\text{epidemic}) \simeq 1-(1/R_0)^{i_0}$ and the probability of no epidemic is, similarly, $\mathbb{P}(\text{no epidemic}) \simeq (1/R_0)^{i_0}$. The main differences that emerge from this stochastic model compared to the deterministic one are the possibility that $R_0 > 1$ does not lead to an outbreak and the final size of the epidemic is given by a distribution rather than a single value.

2.4.1.5 Network based models

In the previous sections, we have been assuming that any individual may infect any other, implying the existence of a fully connected contact network between the individuals in the model. This is an unrealistic assumption, real-world data of contact or social networks shows that these network have properties such as scale-free (high variability in the number of contacts each individual makes), homophily (individuals seek to connect with other individuals with similar properties as themselves) and local clustering effects.

To address this weakness, models based on network structure have been developed, by representing individuals as nodes in a network and edges connecting them as the contact patterns between the individuals, vehicles for the spread of infection. Transmission from an infected node to a susceptible across an edge occurs as a Poisson process with rate τ (the transmission rate, that may vary per edge, which we represent by τ_{ij}). As the transmission across multiple edges is considered to be independent, a susceptible node with k infectious contacts becomes infected according to a Poisson process with rate $k\tau$. An infected node recovers independently of the remainder of the network following a Poisson process with rate γ (the recovery rate), that may change from node to node (in which case we represent as γ_i for node i).

We now show one way to describe an SIS model with network structure. Let $\langle S_i \rangle(t)$

denote the probability that node i is susceptible at time t and $\langle I_i\rangle(t)$ denote the probability that node i is infected at time t. Further, consider the following notation $\langle S_iI_j\rangle(t)$ to denote the probability that node i is susceptible and node j is infected at time t. Also let $g_{ij}=1$ represent the existence of an edge between nodes i and j. The equations that describe an SIS model on a graph can then be written as:

$$
\begin{aligned}
\frac{d\langle I_i\rangle(t)}{dt} &= \tau\sum_{j=1}^{N} g_{ij}\langle S_iI_j\rangle(t) - \gamma_i\langle I_i\rangle(t),\\
\frac{d\langle S_iI_j\rangle(t)}{dt} &= \tau\sum_{k=1,k\neq i}^{N} g_{jk}\langle S_iS_jI_k\rangle(t) - \tau\sum_{k=1,k\neq j}^{N} g_{ik}\langle I_kS_iI_j\rangle(t)\\
&\quad - \tau g_{ij}\langle S_iI_j\rangle(t) - \gamma_j\langle S_iI_j\rangle(t) + \gamma_i\langle I_iI_j\rangle(t),\\
\frac{d\langle I_iI_j\rangle(t)}{dt} &= \tau\sum_{k=1,k\neq i}^{N} g_{jk}\langle I_iS_jI_k\rangle(t) + \tau\sum_{k=1,k\neq j}^{N} g_{ik}\langle I_kS_iI_j\rangle(t) - (\gamma_i+\gamma_j)\langle I_iI_j\rangle(t)\\
&\quad + \tau g_{ij}\langle S_iI_j\rangle(t) + \tau g_{ji}\langle I_iS_j\rangle(t),\\
\frac{d\langle S_iS_j\rangle(t)}{dt} &= -\tau\sum_{k=1,k\neq j}^{N} g_{ik}\langle I_kS_iS_j\rangle(t) - \tau\sum_{k=1,k\neq i}^{N} g_{jk}\langle S_iS_jI_k\rangle(t),
\end{aligned}
$$

recalling that $\langle S_i\rangle(t) = 1-\langle I_i\rangle(t)$, so we do not need to write an equation for the rate of change of $\langle S_i\rangle(t)$. The terms of each equation are simple to explain, for example, the term $\tau\sum_{j=1}^{N} g_{ij}\langle S_iI_j\rangle(t)$ means that the probability of node i being infected in time t depends on the probability that the node is susceptible and the number of its neighbours that are infected, times the transmission rate. Refer to the book by Kiss et al. [83] for an illustrative example in a toy networks of 3 nodes for more intuition on the origin of these equations.

2.4.1.6 Stochastic simulation

In order to solve system of equations as the presented above, stochastic simulation is often used. To this end, the Gillespie algorithm [49, 50] is one of the most widely employed algorithms. Without going into great detail about the procedure, the algorithm proceeds in the following iterative steps:

- Find the rate at which each infected node will recover and each susceptible node may become infected and compute the total rate of change occurring,

- Draw the waiting time until the next event from an exponential distribution parametrized by this total rate of change,
- Select which event occurs, the probability of choosing an event is proportional to the event's rate (for example, if an infection event has a rate of β and a recovery event has a rate of γ, such that $\beta > \gamma$, then the probability of the infection event is higher than the recovery event),
- Update the rates based on the change of status of the nodes involved in the event and repeat.

2.4.2 Agent-based models

Advances in data generation and computing power have led to the emergence of agent-based models, focused on the interactions and behaviours of individuals within complex systems. In ABMs, each agent has a set of defined simple rules, from which emergent complex behaviour is observed at a macro level, thus it is said that ABMs are a bottom-up approach to modelling [99].

There has been widespread adoption of ABMs in infectious disease modelling, with approaches that vary based on the application of the model [172]. As such, the literature is filled with examples of agent-based simulators for epidemiology (see, for example, STDSIM [119], EpiSimdemics [10], FRED [56] or the systematic review by Willem et al. [183]) that have been applied to numerous infectious diseases, such as influenza [35], measles [96, 129] or Ebola [160].

In the context of COVID-19, ABMs have also been widely employed to inform government policy in regards to the impact of policy on disease transmission [124, 125]. Besides Covasim [82], other examples include a model by Wolfram [186], OpenABM-Covid19 [63] or COVID-ABS [162].

Chapter 3

Directed Network Comparison

3.1 Introduction

Undirected networks represent relationships between entities under the assumption that the relationship is symmetrical. For example, a Facebook friendship is mutual because both users mutually agree to become friends and thus there is no difference between source and target in the friendship. This abstraction is often insufficient to capture more nuanced relationships. For instance, in Twitter there is an asymmetry in the follower-followed relationship because the action of following another user is not necessarily reciprocated. Directed networks allow us to model such relationships, but this increased wealth of information leads to greater complexity when analyzing them. Such is the case when comparing directed networks, where network comparison metrics designed for undirected networks have been shown to be unsuitable for the task [7, 152].

The first contribution in this chapter is an extension of NetEmd to directed networks, using directed graphlets of size up to 4. As exposed in Chapter 2, Xu and Reinert [190] have previously proposed a similar extension (named *TriadEMD*) but it is limited to graphlets of size 3. The main obstacle when going from size 3 to size 4 graphlets in directed networks is the combinatorial explosion in number of orbits due to edge direction, there are 33 orbits in graphlets of size up to 3 but 730 when including size 4 as well. In the undirected case, scaling up from size 3 to size 4 is a comparatively much smaller jump, from 4 to 15. We propose two methodologies for our extension to size 4: the first is a simple extension, where all orbits are used in the NetEmd calculation; in the second, we adapt the idea of Aparício et al. [7] and use only orbits that have non zero frequency in both networks for the comparison. We test systematically the

performance difference between using size 3 and size 4 graphlets to understand under which circumstances the added computational load of size 4 graphlets leads to gains in comparison performance. It is also worth noting that other related methods that use graphlets for comparing directed networks, namely Directed Graphlet Correlation Degree (DGCD) [152] and Graphlet Degree Distribution Agreement (GDA) [7] (refer to Section 2.2 for an overview of these methods), also use size 4 graphlets as input to their comparison measures.

One more key difference between our work and *TriadEMD* is the inclusion of orbits from size 2 graphlets in the comparison, which is tantamount to including a comparison of the degree distribution in the network comparison measure. Degree distributions are the most commonly studied property to measure the structure of networks and are able to distinguish between networks created with different models [116, 133, 136], but often they are insufficient as an heuristic for network comparison as it is possible to craft networks with the same degree distribution but widely different structures [94, 133]. Finally, the inclusion of orbits from size 2 graphlets is also supported by the undirected version of NetEmd, where these orbits are also included [180].

Complex networks are heterogeneous in the amount of data they represent, with famous examples ranging from Zachary's karate club [198] with only 34 individuals to gigantic networks like the Friendster social network [191], which contains more than one billion relationships between its users. Such disparity makes the task of comparing networks even harder, as networks that differ in size by multiple orders of magnitude may still be organized according to similar topography, for instance when they are generated by the same process. Identifying such "common organizational principles" [180] is the basis for the NetEmd comparison measure. However, networks representing real world phenomena are often inaccurate representations of that phenomena due to unseen data, measurement errors or simply because the systems they represent are too complex to fully describe. These inaccuracies lead to noise in the statistics used to describe the network and similar topographies become harder to recognize, especially when networks differ greatly in size.

The distributions of graphlets are not immune to these errors, so our second contribution in this chapter is using denoising methods to reduce the impact of these errors so that NetEmd is able to more accurately distinguish networks according to those core structures that compose each network. To this end, we propose a framework that employs Principal Component Analysis (PCA) [68] as a denoising method before comparing the graphlet distributions with Earth Mover's Distance (EMD). PCA assumes

that mean and variance are sufficient statistics of the distribution [159], but graphlet degree distributions break this assumption as, although hard to characterize, they are thought to be similar to degree distributions that approximately follow power law distributions in real world networks [9]. Therefore, we hypothesize that PCA struggles to create an appropriate model for these distributions, which can make the denoising process ineffective, so we also propose an alternative intermediate step that uses Independent Component Analysis (ICA) [25] instead of PCA.

We test our contributions in clustering tasks involving synthetic and real world networks, following the same experimental procedure as the original NetEmd paper [180], which adheres to the framework of graph comparison introduced by Yaveroğlu et al. [193]. We find that our proposed method of denoising using component analysis techniques is able to improve NetEmd's performance on clustering tasks, both in undirected and directed networks. This improvement is particularly significant when comparing networks of different sizes and densities, where NetEmd is already the state of the art.

3.2 NetEmd with Directed Networks

Directed networks pose challenges unlike the ones observed in undirected networks, due to the combinatorial explosion in number of orbits introduced by distinguishing (u,v) and (v,u) as different edges between u and v. There are 730 orbits when considering directed graphlets of size up to 4, compared to 15 in the undirected case, and scaling up to size 5 in directed networks becomes unfeasible as the number of orbits rises to $45,637$. This sharp increase makes the task of counting orbit frequencies for each node even harder, with no known analytical approaches like ORCA [65] that rely on crafting sets of equations that exploit combinatorial relationships between smaller graphlets to compute orbit counts. Instead, enumeration based approaches are required in the directed case, which are known to be at least an order of magnitude slower than analytical approaches in undirected networks. In order to adapt NetEmd to directed networks, we use the G-Trie [140] data structure and the counting algorithm proposed by Aparício et al. [7], publicly available at [138], modified to return the graphlet degree matrix instead of the graphlet degree distribution.

How to handle this increased number of orbits in the comparison is also a challenging problem, especially when comparing networks that contain only a small subset of orbits and differences within those orbits can get diluted when taking the average over the whole set. Aparício et al. [7] argue that the comparison should only be done over

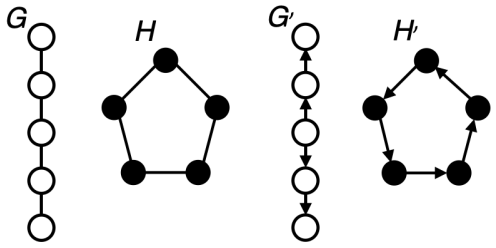


Figure 3.1: Example to illustrate the usefulness of only using orbits that appear in at least one of the networks being compared. Graphs G and H are the example provided by Aparício et al. [7]; we adapt them to directed networks G' and H', keeping the same number of orbits in each network. We measure $NetEmd(G',H') = 5 \times 10^{-5}$ and $Weighted\_NetEmd(G',H') = 0.12$.

the orbits present in at least one of the networks, using networks G and H in Figure 3.1 as an example. In this case, the authors show that using the original formulation of the GDA (described in Section 2.2.3) inflates the similarity score between the two networks far beyond what would be expected from two so distinct looking networks, with an agreement score of 0.92 against 0.32 when using their modified version that compares only orbits in at least one of the networks. This modification to GDA is well-founded because GDA is a measure of *agreement*, bounded between 0 and 1 and meant to be interpretable; a score close to 1 means that the networks are similar and a score close to 0 the opposite, so the magnitude of the value that the measure outputs is of interest. However, NetEmd is a measure of distance between networks and the relative difference between distances is more informative than the absolute value of that distance, so including information that an orbit is missing in both networks improves our ability to tell if they are less distant from each other when comparing against another pair of networks.

With the above in mind, we propose two versions of directed NetEmd. Given two networks G and H and a set of m orbits $O = \{o_1, o_2, \ldots, o_m\}$, the first uses all m orbits in this set, which in practice are all the orbits in graphlets of size up to 3 or 4 and the formulation is the same as the original NetEmd in Equation 2.1. The second is using the same idea of Aparício et al. [7], which we refer to as *weighted NetEmd*, where we restrict the set of orbits to those that occur in at least one of the networks G or H. This leads to a new set of m' orbits $O' = \{o'_1, o'_2, \ldots, o'_m\}$, with $m' \leq m$. The average in Equation 2.1 is done over this set O' instead of the original O, but the rest of the

formula is the same:

$$Weighted\_NetEmd_{O'}(G,H) = \frac{1}{m'}\sum_{i=1}^{m'} EMD^*(p_{o_i}(G), p_{o_i}(H)). \tag{3.1}$$

Graphs G' and H' in Figure 3.1 show an analogous example to the one shown by Aparício et al. [7] to demonstrate the difference between weighted and original NetEmd applied to directed networks. Although it is usually impractical to scale up to size 5 in directed networks, in a small example like this one it becomes feasible to do so. In this context, we calculate a distance of 5×10^{-5} using all orbits and 0.12 when using the weighted version.

3.3 NetEmd with Dimension Reduction

3.3.1 Motivation

In this section, we describe how we couple dimensionality reduction techniques with NetEmd, using them as noise reduction techniques. The rationale is that processes that generate networks are inherently noisy due to the complexity of the systems they represent. This is particularly evident when we consider random networks generated by the same model; we expect them to have similar structure and that similarity to be reflected in the distribution of subgraph and orbit frequencies, but the stochastic nature of the generation process introduces noise in these distributions that makes them harder to identify.

Considering the graphlets and orbits shown in Figure 2.1, it is clear that (for example) directed graphlet G_2 cannot exist if G_0 does not, and similarly directed graphlet G_{12} cannot exist if G_1 does not. More generally, it is clear that the different orbits do not represent independent degrees of freedom, and are expected to be constrained by the requirement of combinatorial consistency of the underlying full graph G as well as correlations induced by the generation process.

In practice, both combinatorial enumeration of graphlets and *a priori* determination of statistical relationships between them (beyond comparison to simple random graphs as in [108]) are computationally infeasible. While this does not matter in practice for the undirected case, for directed graphs as we consider here, the relatively large number and complexity of orbits makes these much more important. We therefore seek an approach that can adjust for the non-independence of orbit counts.

Proceeding somewhat formally, let $\mathbf{F}_G$ denote the graphlet degree matrix of graph G for a set of orbits O, a $n \times m$ matrix where $n = |V(G)|$, $m = |O|$, such that $[\mathbf{F}_G]_{i,j}$ represent the frequency of orbit j for node i. Further, let $\mathbf{f}_i$ be the vector of frequencies of node i, i.e. $(\mathbf{f}_i)_j = [\mathbf{F}_G]_{i,j}$ meaning $\mathbf{f}_i$ is the i-th row of $\mathbf{F}_G$. Now note that we are only interested into networks up to isomorphism; in fact, our distance measures depend on empirical histograms of orbit counts, deriving the distribution p_{o_i} in (2.1) from histogram heights like

$$h_{o_j,y} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{(\mathbf{f}_i)_j = y\}}, \tag{3.2}$$

which is the proportion of nodes with y counts of orbit o_j, where $\mathbf{1}$ is an indicator function taking the value 1 if its argument is true and the value 0 otherwise. The right-hand side of (3.2) is a sum of random variables, one for each node, and from (2.1) we will further sum over orbits for a function of two such histogram heights.

In general, we will expect a random graph model to assign a probability to each graph in some finite set, and according to this measure we expect a probability distribution to be induced on $\mathbf{h}_G = [h_{o_j,y}]$. As discussed above, we expect both combinatorial constraints and correlations between orbit counts, meaning that a sum over all elements of $\mathbf{h}_G$ as implied by (3.2) and (2.1) will inflate, due to the additivity of variances of random variables, the amount of noise in EMD realisations compared to the amount that is absolutely necessary under the random graph model.

Therefore, a natural approach is to denoise using techniques that do not require explicit solution of or simulation from random graph models. In particular, we will use the linear techniques of principal component analysis (PCA) and independent component analysis (ICA), noting their computational efficiency compared to potentially more general nonlinear methods [64]. The main idea of our methodology is therefore to project the orbit frequencies to a lower dimension, training the dimension reduction model to lose minimal information while removing noise.

Since it is not guaranteed that each graph will dimensionally reduce to the same size, the approaches we use allow for expansion of the dimensionally reduced features back to the original (high) dimension. The NetEmd comparison from Equation 2.1 is consequently applied to the reconstructed frequencies. By using the first L components of the linear methods, which contain the most signal from the dataset, to reconstruct the original counts, this can be intuitively understood as decreasing noise within the orbit frequencies when we do not have a good model for that noise.

3.3.2 Principal Component Analysis

Principal Component Analysis (PCA) is a technique for dimensionality reduction that preserves as much variability of the data as possible, by computing principal components of the data. Principal components are sequences of orthogonal unit vectors that form an orthonormal basis, in which the original dimensions of the data are linearly uncorrelated. When projecting the data onto this new basis, these vectors can be seen as the directions that maximize the variance of the projected data, with the first principal component representing the maximum variance.

Let $\mathbf{F}_G$ denote the graphlet degree matrix of graph G for a set of orbits O, a $n \times m$ matrix where $n = |V(G)|$, $m = |O|$ and $\mathbf{F}_{G_{i,j}}$ represents the frequency of orbit j for node i. We assume that the frequencies have been normalized, a preprocessing step advised before applying PCA (to prevent notation overload, we use the same $\mathbf{F}_G$ to denote the normalized version of the graphlet degree matrix). The principal components of this matrix are defined by $\mathbf{V} = \mathbf{F}_G\mathbf{W}$, where $\mathbf{W}$ is a $m \times m$ matrix whose columns are the eigenvectors of $\mathbf{F}_G^\intercal\mathbf{F}_G$. To allow for only the first L components, the matrix $\mathbf{W}$ can be truncated to $\mathbf{W}_L$, with dimensions $m \times L$, leading to a transformation $\mathbf{V}_L = \mathbf{F}_G\mathbf{W}_L$. This truncation is done such that the L eigenvectors chosen to correspond to the largest eigenvalues of $\mathbf{W}$. The original graphlet degree matrix can be reconstructed as $\hat{\mathbf{F}}_G = \mathbf{V}_L\mathbf{W}_L^\intercal = \mathbf{F}_G\mathbf{W}_L\mathbf{W}_L^\intercal$. The goal of PCA is to learn $\mathbf{W}_L$ such that the variance of the original data preserved is maximized, while also minimizing the total squared reconstruction error $||\mathbf{F}_G - \hat{\mathbf{F}}_G||_2^2 = ||\mathbf{V}\mathbf{W}^\intercal - \mathbf{V}_L\mathbf{W}_L^\intercal||_2^2$.

We use the reconstructed graphlet degree matrix to compare networks with NetEmd, keeping Equation 2.1 virtually unchanged from the original formulation:

$$PCA\_NetEmd_O(G,H) = \frac{1}{m}\sum_{i=1}^{m} EMD^*(\hat{p}_{o_i}(G), \hat{p}_{o_i}(H)), \tag{3.3}$$

where $\hat{p}_i(G)$ and $\hat{p}_i(H)$ are obtained from $\hat{\mathbf{F}}_G$ and $\hat{\mathbf{F}}_H$ respectively, instead of $\mathbf{F}_G$ and $\mathbf{F}_H$.

3.3.2.1 Choosing the number of components.

Choosing an appropriate number of components to project the data down affects the reconstruction error and therefore the amount of noise reduced. If we take $L = m$, then clearly we are able to reconstruct the original frequencies perfectly and no noise has been removed from the data. On the other hand, picking a value for L too low, leads to foregoing descriptiveness of the data. A common strategy for picking the number

of components is through the amount of variance explained by each component, a strategy that allows us to adapt the number of components to suit the networks we are comparing.

The sample covariance matrix of each orbit, i.e. $F_{G_{i,1}}, F_{G_{i,2}}, \ldots, F_{G_{i,m}}$, is proportional to $\mathbf{F}_G^{\mathsf{T}}\mathbf{F}_G$ (the derivation for this result can be found in [79], pp. 30-31) and contains the variance of the frequency of each orbit in its diagonal. As a square matrix, according to the spectral theorem, the covariance matrix can be diagonalized by its eigenvectors and the values of the resulting diagonal matrix are the corresponding eigenvalues. These eigenvalues represent variability of each axis in the projected space. Therefore, by taking the sum of the L highest eigenvalues, the ones corresponding to the first L components, we get the variance explained by the first L principal components. The ratio $\sum_1^L \lambda_i / \sum_1^m \lambda_i$, where λ_i is the eigenvalue corresponding to the ith eigenvector, measures the proportion of variance explained by the first L principal components. Finally, to condition the number of components on the proportion of variance explained, the smallest L is calculated such that $\sum_1^L \lambda_i / \sum_1^m \lambda_i \geq r$ where $r \times 100\%$ is the percentage of variance explained.

3.3.3 Independent Component Analysis

Independent Component Analysis (ICA) is a statistical technique in which an observed vector of random variables is thought to be the linear combination of unknown independent components. These components are assumed to be mutually statistically independent and with zero mean. The classic example of an ICA application is a party where a microphone is picking up multiple conversations and the goal is to separate the source signal from each conversation from the mixed one picked up by the microphone.

Transposing ICA to the domain of network comparison, we can draw an analogy from the above application by considering that the voices in the conversation are the nodes in the network and the orbit frequencies are the data points captured by the microphone. Our goal is then to search for the source signals, i.e., network characteristics stemming from its generation mechanism that generate such frequency distribution. We conjecture that by using these source signals to reduce noise in the orbit frequency distributions, our network comparison measure becomes able to more accurately distinguish networks with different generation mechanisms.

As before, let $\mathbf{F}_G$ denote the graphlet degree matrix of graph G for a set of orbits O and let $\mathbf{f}_i$ be the vector of frequencies of node i, i.e. $f_{i,j} = \mathbf{F}_{G_{i,j}}$. We can write $\mathbf{f}_i$

as $\mathbf{f}_i = \mathbf{A}\mathbf{s}_i$, where $\mathbf{A}$ is called the mixing matrix and $\mathbf{s}_i$ are the L independent components. The goal of ICA is to estimate $\mathbf{A}$ and $\mathbf{s}_i$ from $\mathbf{f}_i$ only. In practice, algorithms to calculate independent components compute a weight matrix $\mathbf{W}$, a pseudo-inverse of $\mathbf{A}$ and obtain the independent components through $\mathbf{s}_i = \mathbf{W}\mathbf{f}_i$. Using the FastICA algorithm [72, 73], $\mathbf{W}$ is constructed iteratively by finding unit vectors $\mathbf{w}$ such that the projection $\mathbf{w}^\intercal\mathbf{f}_i$ maximizes non-gaussianity, measured by the approximation of negentropy $J(\mathbf{w}^\intercal\mathbf{f}_i) \propto [E(G(\mathbf{w}^\intercal\mathbf{f}_i)) - E(G(\nu))]^2$, where $G(x) = \log\cosh x$ and ν is a standard Gaussian random variable with mean 0 and variance 1. These unit vectors $\mathbf{w}$ are combined into the weight matrix $\mathbf{W}$ and decorrelated to prevent convergence to the same maxima by $\mathbf{W} = \mathbf{W}/\sqrt{\|\mathbf{W}\mathbf{W}^\intercal\|}$ (the matrix norm is calculated using the l^2 norm); this procedure is then repeated with $\mathbf{W} = \frac{3}{2}\mathbf{W} - \frac{1}{2}\mathbf{W}\mathbf{W}^\intercal\mathbf{W}$ until convergence.

Upon obtaining the weight matrix, we calculate its pseudo-inverse to obtain the mixing matrix $\mathbf{A}$ and calculate $\hat{\mathbf{F}}_G$ similarly to what was done with PCA: we project $\mathbf{F}_G$ to a lower dimensional space using $\mathbf{W}$ and apply the mixing matrix $\mathbf{A}$ to retrieve the original data. The final NetEmd formula then takes a similar form as for *PCA_NetEmd*:

$$ICA\_NetEmd_O(G,H) = \frac{1}{m}\sum_{i=1}^{m} EMD^*(\hat{p}_{o_i}(G), \hat{p}_{o_i}(H)), \tag{3.4}$$

with $\hat{p}_i(G)$ and $\hat{p}_i(H)$ obtained from $\hat{\mathbf{F}}_G$ and $\hat{\mathbf{F}}_H$ respectively, as before.

3.4 Experimental Setup

3.4.1 Measure of Cluster Performance

Given a set of networks $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$, divided in disjoint classes $C = \{c_1, c_2, \ldots, c_m\}$, we use the performance measure proposed by Wegner et al. [180] that captures the idea that networks of the same class should be nearer to each other than to networks of other classes. It is defined as an empirical probability $P(G) = d(G,G_1) < d(G,G_2)$, where G_1 is a network selected randomly from the same class as G and G_2 is randomly selected from a different class, and d is the network comparison statistic. The performance over the whole dataset is the average $P(G)$ over all the networks in $\mathcal{G}$, and we write this as $\overline{P} = P(\mathcal{G}) = \frac{1}{|\mathcal{G}|}\sum_{G\in\mathcal{G}} P(G)$. In Appendix A, we present results using area under precision-recall curve (AUPR) and adjusted Rand index (ARI).

3.4.2 Synthetic datasets

We reproduce the experimental setup of Wegner et al. [180], testing our proposed modifications to NetEmd on the same eight random network models. These models are:

- Erdős-Rényi model [38]. A random graph with n is generated by picking m unique edges chosen at random from the $\binom{n}{2}$ possible edges. m was chosen to generate networks with the appropriate density.

- Barabasi-Albert preferential attachment model [9]. An initial graph is created with m nodes and new nodes are added iteratively to the network, each new node is connected to m existing nodes, picked randomly with probability proportional to their degree. m was chosen to generate networks with the appropriate density.

- Geometric random graphs [48], using a 3-dimensional square ($D = 3$). Nodes are randomly embedded in a D-dimensional space and are connected if the Euclidean distance between them is smaller than a threshold r. This threshold parameter was determined by grid search to generate networks with the desired number of edges.

- Geometric gene duplication model [62]. Starting with an initial network of 5 nodes embedded in a 3-dimensional space, on each iteration a node is chosen randomly to be duplicated. The duplicated node is placed randomly in the 3-dimensional space at an Euclidean distance of at most 2 from the original node. This process is repeated until the desired number of nodes, and nodes at a distance of r or less at connected. This distance r is chosen to generate a network with the appropriate number of edges.

- Duplication divergence model of Vázquez et al. [176]. The network grows in two stages. In the first stage, a node v is chosen randomly to be duplicated into a node v', that keeps all the edges from v. The nodes v and v' are connected with probability 0.05. In the second stage, one of the duplicated edges ((v,u) or (v',u)) is chosen randomly and deleted with probability q. This process is repeated until the network grows to the desired number of nodes and q is chosen to generate networks with the appropriate density.

- Duplication divergence model of Ispolatov et al. [74]. Starting from a network with a single edge, a node is chosen randomly and duplicated, with the new node

keeping each of the original's neighbours with probability p. This parameter is chosen to generate networks with the desired number of edges.

- Configuration model, using the Duplication divergence model of Vazquez et al. [176] as the generator for the graphical degree sequence.

- Watts-Strogatz model [179]. Nodes are placed in a ring and connected to their k nearest neighbours in both sides of the ring. Each edge is then rewired with probability 0.05 to a new node selected at random. The parameter q is chosen to generate networks with the appropriate average degree.

We generate 10 networks per model per combination of number of nodes and average degree. The set of number of nodes used is $N = \{1250, 2500, 5000, 10000\}$ and the set of average degrees is $k = \{10, 20, 40, 80\}$. This leads to a total of 160 networks per model and 1280 networks in total.

For the directed datasets, we add varying levels of reciprocity $\{0, 25, 50, 75, 100\}$. To generate a directed network with reciprocity r, we take the undirected version, we duplicate and invert all edges (if (u, v) is in the undirected network, we add (v, u) to the directed version) and we then take $(100 - r)\%$ of these pairs and choose randomly a direction to remove (either (u, v) or (v, u)). This is done incrementally such that if (u, v) is removed from the dataset with 75% reciprocity, then it is also removed from the datasets with 50%, 25% and 0% reciprocity.

These datasets of synthetic networks give rise to two tasks aimed at gauging how well a network comparison measure separates clusters according to their generation mechanism, with the ground truth given by the random network model used to generate the networks in each cluster. The first, more simple task is to separate networks with the same number of nodes and the same density. To this end, we create 16 groups based on the combinations of N and k, each group containing 10 realizations of each random network model, for a total of 80 networks. This task is equivalent to RG_1 in the original NetEmd paper [180]. The second task, equivalent to RG_3 in the original NetEmd paper, is to compare all 1280 networks simultaneously, finding the 8 clusters of 160 networks. This task measures how sensitive the comparison measure is to differences in an order of magnitude in number of nodes and edges, making the separation by model type more difficult. In directed networks, we repeat Task 1 and 2 for each level of reciprocity, determining how sensitive the network comparison is to this third parameter that impacts the set of orbits available for comparison (the set of orbits with 0% reciprocity is disjoint from the set of orbits with 100% reciprocity).

		Onnela et al.	Real directed
# Networks		151	1283
Number of Nodes	Min.	30	9
	Median	918	160
	Max	11586	62586
Number of Edges	Min.	62	21
	Median	2436	2428
	Max	232794	1614977
Density	Min.	4.26×10^{-5}	3.78×10^{-5}
	Median	0.015	0.095
	Max	0.499	0.75
Reciprocity	Min.	-	0.0
	Median	-	0.533
	Max	-	1.0

Table 3.1: Summary statistics for the Onnela et al. and real world directed networks datasets.

3.4.3 Real world datasets

We use the dataset of Onnela et al. [118] to validate our methodology on a mix of real world and synthetic networks. The multiple sources that make up this dataset lead to a heterogeneous set of networks; however, the ability to separate these networks according to their domain is desired from a network comparison method. We were unable to find sources for the original 746 networks, so we use the reduced set proposed by Ali et al. [4] with 151 unweighted and undirected networks. There is no ground truth for the true clusters in this dataset and the dendograms constructed through the methods proposed by Onnela et al. [118] and Ali et al. [4] disagree on the composition of each cluster. Therefore, we aim to reconstruct clusters according to the type of data, which can be visualized in the supplementary material of Ali et al. [4].

For directed networks, we download four datasets from the Stanford Network Analysis Project (SNAP) library [93]: Gnutella peer-to-peer file sharing network from August 2002 [92, 142] (9 networks with an average reciprocity of 0%), CAIDA autonomous systems relationships datasets from January 2004 to November 2007 [91] (122 networks with an average reciprocity of 100%), ego networks of *circles* from Google+ [105] (132 networks with an average reciprocity of 28%) and ego networks of *lists* from Twitter [105] (973 networks with an average reciprocity of 54%).

The summary statistics of the Onnela et al. [118] and the real world directed networks are given in Table 3.1.

Algorithm	Parameter	Synthetic		Real
		Task 1	Task 2	Onnela et al.
Original NetEmd	G5	*0.995 ± 0.003*	0.918	0.890
PCA_NetEmd	50% Variance	0.98 ± 0.01	0.915	0.864
	80% Variance	0.98 ± 0.01	0.918	**0.893**
	90% Variance	*0.995 ± 0.004*	0.922	0.891
	95% Variance	*0.993 ± 0.004*	0.919	0.892
	99% Variance	*0.995 ± 0.004*	0.921	0.891
ICA_NetEmd	2 components	0.985 ± 0.002	0.926	0.869
	10 components	*0.994 ± 0.004*	**0.930**	0.887
	15 components	**0.996 ± 0.003**	0.929	0.890
GCD	11 orbits	*0.99 ± 0.01*	0.888	0.789
	73 orbits	*0.994 ± 0.008*	0.885	0.819
GDA	G3	0.986 ± 0.003	0.868	0.876
	G4	0.977 ± 0.007	0.886	0.855
	G5	0.95 ± 0.01	-	0.832

Table 3.2: Results for Task 1, Task 2 and the Onnela et al. [118] datasets in undirected networks. The metric used for Task 1 is the mean (and standard error of the mean) of the 16 values for $\overline{P}$ in each combination of number of nodes with network density. The metric used for Task 2 and Onnela et al. dataset is the sole value of $\overline{P}$ after comparing the 1280 and 151 networks, respectively. The parameter column indicates: graphlet size used in original NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 5); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 5; the number of orbits used by Graphlet Correlation Degree (GCD); graphlet sizes used in GDA. The bolded values are the maximum for each task. Italic values are within one standard error of the maximum performance. Note that results for GDA with graphlet size 5 took longer than a week to return results for Task 2, at which point we stopped the computation.

3.5 Results

3.5.1 Undirected Results

We present the results for Tasks 1 and 2 in undirected networks and the results for the Onnela et al. [118] dataset in Table 3.2. The results for Task 1 show the average and standard error for the 16 values of $\overline{P}$, one for each combination of N and k. The results for Task 2 and Onnela et al. dataset show the single value of $\overline{P}$ after comparing the 1280 and 151 networks, respectively.

We find that our proposed modifications to NetEmd achieve an increase in performance, with *ICA_NetEmd* outperforming original NetEmd in synthetic datasets and

PCA_NetEmd in the Onnela et al. dataset. The performance gain in Task 1 is not significant as the results are within a standard error of each other, but on Task 2, the more difficult task, the difference is more pronounced. This performance improvement in synthetic datasets is not reflected in the Onnela et al. dataset, where 15 components are necessary to equalize the performance of original NetEmd. The best performance in the Onnela et al. dataset is *PCA_NetEmd* with 80% explained variance, which contrasts with the results in the synthetic dataset, where 80% explained variance achieves worse performances than the original NetEmd. Using higher values for explained variance in *PCA_NetEmd* also improves performance over original NetEmd in Task 2 and Onnela et al. dataset, but not in Task 1.

3.5.2 Directed Results

We show the results for Task 1 in directed networks in Table B.1 and the results for Task 2 and the dataset of real world directed networks in Table 3.3. The results for Task 1 show the average and standard error for the 16 values of $\overline{P}$, one for each combination of N and k, for each level of reciprocity. The results for Task 2 and real world directed networks dataset show the single value of $\overline{P}$ after comparing the 1280, for each level of reciprocity, and 1232 networks, respectively.

Similarly to the undirected case, we find that the results for each algorithm in Task 1 are similar across different levels of reciprocity, meaning that the various different versions of NetEmd and DGCD are able to distinguish the generation mechanism regardless of how many edges are reciprocated in a directed network. Results show that DGCD with 129 orbits achieves best performance for 0%, 25%, 50% and 75% and *PCA_NetEmd* with 90% explained variance for 100% reciprocity. We observe no significant difference between using NetEmd with all orbits, with weighted orbits or coupled with dimensionality reduction techniques. Results with TriadEMD are non-significantly better than our NetEmd version with orbits from size 3 graphlets in this task, indicating that using graphlets of size 2 for the comparison in this task does not help differentiating between different models. This is likely due to the configuration model, which shares the same degree distribution as the duplication divergence model of Vazquez et al. [176].

In Task 2, our results show a noticeable gain in performance when using NetEmd with ICA over other NetEmd versions and DGCD, in particular when using only 2 components and in reciprocity levels smaller than 100%. We find that using orbits in graphlets of only size 2 and 3 orbits achieves better results for this task than using

Algorithm	Parameter	Synthetic					Real Networks
		Reciprocity					
		0%	25%	50%	75%	100%	
All Orbits	G3D	0.923	0.905	0.907	0.905	0.915	0.864
	G4D	0.908	0.879	0.881	0.878	**0.927**	0.813
Weighted	G3D	0.918	0.899	0.902	0.898	0.913	**0.868**
	G4D	0.902	0.876	0.878	0.875	0.923	0.827
PCA	50% Variance	0.924	0.897	0.900	0.894	0.908	0.821
	80% Variance	0.924	0.885	0.892	0.881	0.920	0.825
	90% Variance	0.921	0.882	0.889	0.878	**0.927**	0.824
	95% Variance	0.915	0.880	0.887	0.876	0.922	0.822
	99% Variance	0.914	0.878	0.885	0.875	0.924	0.818
ICA	2 components	**0.941**	**0.916**	**0.914**	**0.911**	**0.927**	0.712
	10 components	0.934	0.900	0.903	0.896	0.923	0.843
	15 components	0.931	0.896	0.903	0.892	0.917	0.844
TriadEMD		0.926	0.903	0.906	0.906	0.916	0.859
DGCD	13 orbits	0.905	0.904	0.895	0.885	0.871	0.812
	129 orbits	0.887	0.884	0.880	0.880	0.885	0.799
GDA	G3D	0.842	0.802	0.802	0.800	0.858	0.736
	G4D	-	-	-	-	0.886	-

Table 3.3: Results for Task 2 in directed networks, for each level of reciprocity, and for the real world networks dataset. The metric used for this task is the sole value of $\overline{P}$ after comparing the 1280 and 1231 networks, respectively. The parameter column indicates: graphlet size used in directed and weighted NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 4); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 4; the number of orbits used by DGCD; graphlet sizes used in GDA. The bolded values are the maximum for each dataset. Note that results for GDA with graphlet size 4 took longer than a week to return results for Task 2 and in the real world networks dataset, at which point we stopped the computation.

ones with size up to 4, with the exception of the two extremities in reciprocity (0% and 100%), where having more available orbits for comparison helps performance. In this task, we also find that using *Weighted_NetEmd* degrades performance compared to using all orbits. Results also show that our versions of NetEmd perform very similarly TriadEMD, with a slight advantage to TriadEMD at 0, 75% and 100% reciprocity, which is expected again due to the configuration model adding a confounding factor when using graphlets G_0 and G_1 from Figure 2.1 in the comparison.

On the other hand, *Weighted_NetEmd* reaches the best performance in the real directed networks dataset, where, similarly to Task 2 in synthetic networks, using

smaller graphlets yields better results. Table 3.3 shows that *ICA_NetEmd* with 10 and 15 components obtains better performance than NetEmd without dimensionality reduction and with size 4 graphlets. This comparison is unfair to *ICA_NetEmd*, as results suggest that when there is a large discrepancy between the sizes of the networks, using smaller graphlet sizes improves the performance of NetEmd. Therefore, although *Weighted_NetEmd* achieves the best performance out of the parameters we include in Table 3.3, the top performance in the real world directed networks dataset is achieved by *ICA_NetEmd* with size 3 graphlets and 10 components. All four versions of NetEmd that we propose achieve better results than DGCD and GDA in this dataset. This is also true when comparing our versions with size 3 graphlets against TriadEMD, showing that in real world networks the degree distribution is able to contribute in a positive manner towards being able to discriminate networks of different sources

3.5.3 Performance by number of ICA components

We examine the impact on performance of varying the number of components used for *ICA_NetEmd*, with the caveat that increasing the number of components leads to a greater computational complexity as the FastICA algorithm takes more iterations to converge. We also explore the performance of *ICA_NetEmd* under smaller graphlet sizes, namely graphlets sizes 3 and 4 for undirected networks and size 3 for directed. There are two different ways of using smaller graphlets on *ICA_NetEmd*, the more natural way in which we compute the independent components using the set of orbits related to the desired graphlet size, or an alternative way in which the components are computed using a larger graphlet size and the reconstructed graphlet degree matrix is posteriorly truncated to include only the orbits of the desired graphlet size. For example, when calculating network distances using undirected graphlets of size up to 4, there are 15 orbits. In the first method, the graphlet degree matrix $\mathbf{F}_G$ is a $|V(G)| \times 15$ matrix and we simply calculate $\hat{\mathbf{F}}_G$ as described in Section 3.3, constraining the number of components used to a maximum of 14 components. In the second method, $\mathbf{F}_G$ is a $|V(G)| \times 73$ matrix, allowing a maximum of 72 components to compute $\hat{\mathbf{F}}_G$, which is then truncated to 15 columns.

Figures 3.2 to 3.7 show, in order, performance with varying number of components in undirected networks Task 1, undirected networks Task 2, directed networks Task 1, directed networks Task 2, the Onnela et al. dataset and directed real world networks.

We find that adding more components to ICA does not always translate to a better performance, instead we find that there is a performance maximum obtained with a

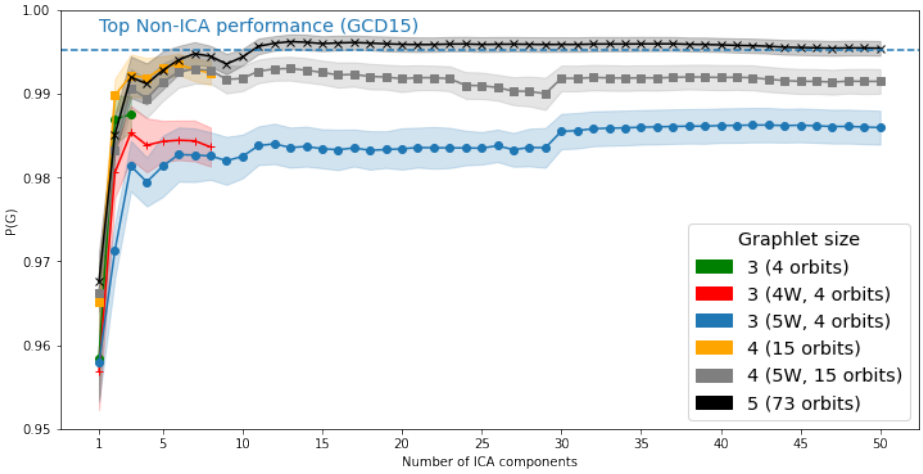


Figure 3.2: Undirected networks - Task 1

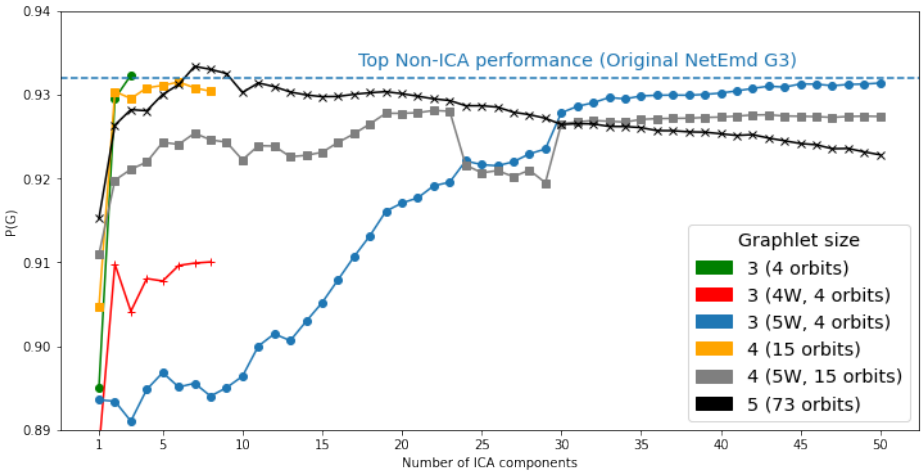


Figure 3.3: Undirected networks - Task 2

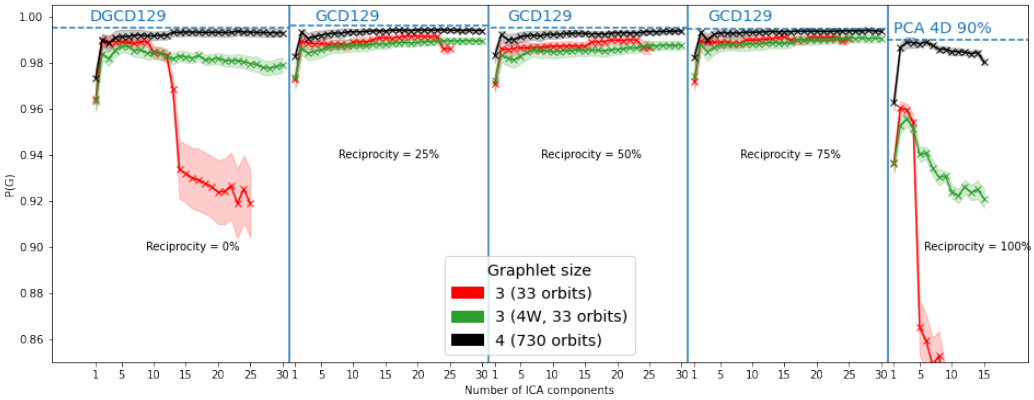


Figure 3.4: Directed networks - Task 1

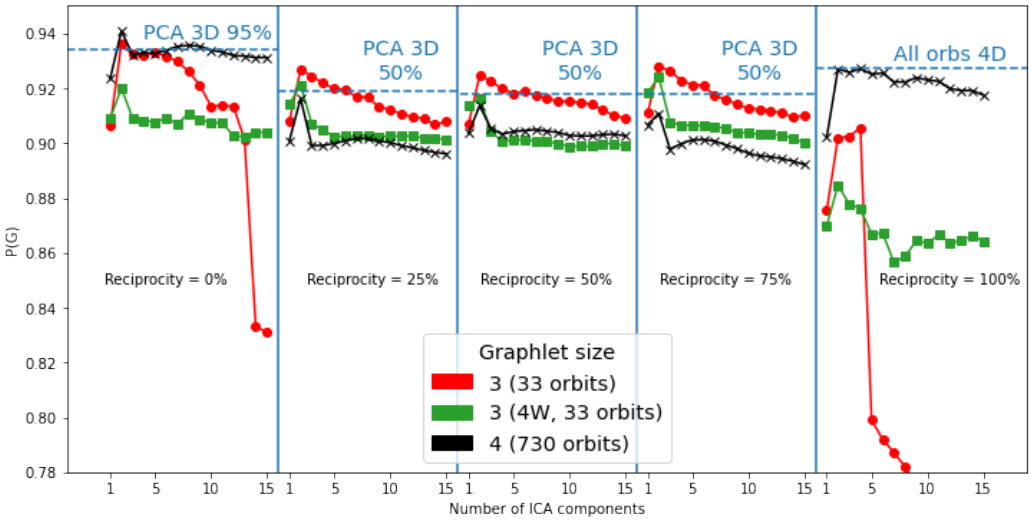


Figure 3.5: Directed networks - Task 2

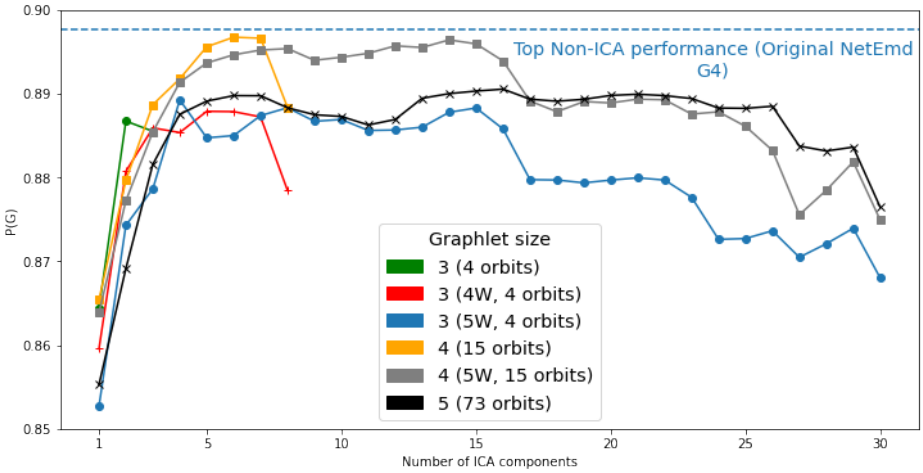


Figure 3.6: Onnela et al. dataset

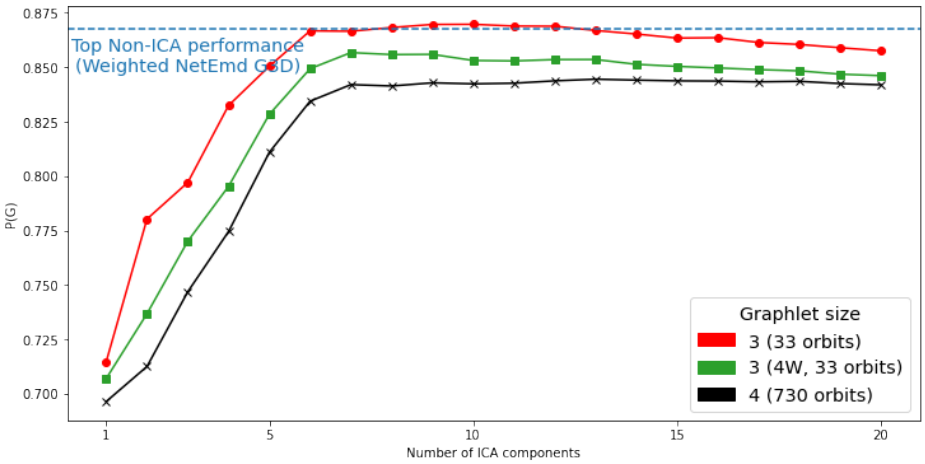


Figure 3.7: Real world directed networks dataset

low number of components that depends on the task. Adding more components after this maximum leads to a decline in performance, either an immediately visible decline (Figures 3.3, 3.5 and 3.6) or after plateauing at that maximum value (Figures 3.2, 3.4 and 3.7). This behaviour indicates that our ICA algorithm is working as expected, as we find a suitable number of components that maximizes the noise reduced in the orbit frequency distributions, so adding more components reduces the difference between $\mathbf{F}_G$ and $\hat{\mathbf{F}}_G$ which negates the positive impact on that noise reduction.

Figures 3.4 and 3.5 show a sharp decline in performance when using graphlets of size up to 3 with reciprocity 0% and 100%, after 13 and 4 components respectively. This is caused by these extreme levels of reciprocity limiting the orbits that can occur in the networks we compare. With 0% reciprocity, only graphlets with no reciprocal edges can occur in these networks, so all orbits in graphlets with reciprocal edges have a frequency of 0. As there are only 13 orbits belonging to graphlets with no reciprocal edges, when we try to compute more than 13 independent components the performance degrades as more components are added. With 100% reciprocity, we find a similar situation but there are only 4 orbits in graphlets of size 2 and 3 for which all edges are reciprocal (orbits 2, 27, 28 and 32 in Figure 2.1).

3.5.4 Discussion on the time complexity of NetEmd

Calculating the NetEmd measure is a process that can be separated in two phases: acquiring the distributions of the network statistics of interest and comparing the statistics using the EMD. In the case of the algorithm we propose in this work, obtaining the distributions of network statistics involves calculating the graphlet degree matrix and performing PCA or ICA on this matrix.

Obtaining the graphlet degree matrix is the most computationally expensive step of this process, with a time complexity of $O(Nd^{m-1})$, where N is the number of nodes in the network, d the maximum degree of any node in the network and m the size of the graphlets being enumerated. In undirected networks, we use the combinatorial algorithm *ORCA* [65] that relies on an analytical approach to set up a system of linear equations that relates different orbit frequencies. In directed networks, no such approaches are known [139], so we rely on *G-Tries* [7, 140], a data structure that supports a graphlet enumeration algorithm by representing subgraphs in a prefix tree, the state of the art in enumerating directed graphlets, to our knowledge [7, 139].

The complexity of PCA is split in two parts, computing the covariance matrix is $O(Np^2)$ and doing the eigenvalue decomposition is $O(p^3)$, where N is the number of

nodes in the network and p the number of orbits being considered. This leads to an overall complexity of $O(Np^2 + p^3)$.

The complexity of ICA is harder to characterize. Firstly, the FastICA algorithm assumes that the data has been centered and whitened [73]. The implementation we use, from scikit-learn [127], uses PCA to do the whitening preprocessing, so its complexity is at least as high as PCA. The iterative algorithm to find the weight matrix $\mathbf{W}$ is repeated until the matrix has converged, but there is no prior assurance that the algorithm converges or how many iterations it takes to do so. The scikit-learn implementation defines a maximum number of iterations to stop execution, we set this value to 1000. Each iteration involves calculating $\log\cosh(\mathbf{w}^\intercal \mathbf{f}_i)$ for each component and each node, which the scikit-learn implementation does in $O(Nc^2)$, where c is the number of components to be calculated, and the decorrelation step, which is done in $O(c^2)$. The overall complexity of ICA then becomes $O(Np^2 + p^3 + INc^2)$, where I is the number of iterations.

Wegner et al. [180] calculate the complexity of comparing two graphlet distributions using EMD^* to be $O(k(N+N')\log(N+N'))$, where N and N' are the number of nodes of each network and k is the maximum number of function calls to the optimization algorithm used to align the distributions.

3.5.5 Discussion on orbits from smaller graphlet sizes

Wegner et al. [180] have shown that reducing the graphlet size as input to NetEmd can improve the quality of the clusters under certain conditions. A more common reason to use smaller graphlets for comparison is a computational load argument, as we show in Appendix 3.5.4, increasing the graphlet size leads to more computation time in each step of the NetEmd framework. In the first step, computing orbit frequencies of graphlets of size 4 in directed networks or size 5 in undirected networks can be prohibitive for very large graphs. Another consideration is the time taken to compute the EMD between orbit distributions, particularly in directed networks with graphlets of size 4 requiring 22 times more calls to the EMD^* function than using size 3. Finally, computing the principal or independent components also contain a complexity component dependent on the number of orbits we consider as features to these methods, which also contribute to greater computational load when increasing the graphlet sizes as input to NetEmd. Therefore, it is important to understand in which situations decreasing the graphlet size as input to NetEmd leads to similar or better performance to avoid paying the high computational cost.

We previously presented the performance of *ICA_NetEmd* under smaller graphlet sizes, for multiple components used, in Figures 3.2 to 3.7. The results for Task 1, both in directed and undirected networks, demonstrate that when comparing networks of similar sizes and densities using bigger graphlet sizes yields a more accurate comparison. The same holds for Task 2 in the undirected case and in the extreme cases of reciprocity (0% and 100%) in Task 2 of the directed experiments, where ICA achieves the top performance with size 5 undirected and size 4 directed graphlets, respectively. On the other hand, in the datasets of real world networks and in directed Task 2 with reciprocity of 25%, 50% and 75%, size 5 graphlets in undirected networks and size 4 in directed networks perform worse than size 4 and size 3, respectively.

For the datasets of real networks in particular, this difference in performance can be explained by noticing that these datasets contain very small networks such that some graphlets have no matches in them. This issue is amplified the larger the graphlet size considered is and it adds a confounding factor to the network comparison, similar to the example of Figure 3.1. Incidentally, this is the only dataset where *Weighted_NetEmd* achieves better performance than following the original formulation.

Figures 3.2 to 3.7 also show the performance results of using the full set of size k orbits to denoise the distributions of $k-1$ or $k-2$ orbits. In this second way of using ICA, the only time saved compared to using the full set of orbits is the time to compute the EMD between orbit distributions and we find that in no case it leads to the best clustering performance. However, Figure 3.3 shows a breakpoint at 30 components, where the performance using size 3 and 4 orbits after denoising with size 5 succeeds in performing better than size 5. The reason for this is that, as more components are added, the smaller the reconstruction error $||\mathbf{F}_G - \hat{\mathbf{F}}_G||_2$ is, so the performance with those parameters tends to approximate the performance of the original NetEmd with those graphlets sizes.

The performance of *PCA_NetEmd* under smaller graphlet sizes follows a similar set of rules to *ICA_NetEmd*, with better performances in Task 1 measured when using larger graphlets and in Task 2 and real world networks when using smaller graphlet sizes. Tables B.2 to B.4 detail the results for multiple combinations of graphlet size and percentage of variance explained. Unlike *ICA_NetEmd*, our experimental setup does not allow us to claim that after a certain value of explained variance performance degrades or stabilizes. Instead, results suggest that the optimal threshold depends on both the dataset and the size of graphlet chosen. Like with *ICA_NetEmd*, choosing a low value for percentage of explained variance for Task 2 in directed networks yields

the best performance for 0, 25, 50 and 75% reciprocity, but otherwise choosing 90 or 95% explained variance seems to be a good rule of thumb for best performance with *PCA_NetEmd*.

3.6 Conclusion

We presented two extensions of NetEmd to directed networks, one a direct extension that uses all orbits of graphlets up to size 4 and another that compares networks based only on the orbits that exist in at least one of the networks. We showed that our methodology achieves state of the art performance in large datasets of synthetic networks with heterogeneous network sizes and in datasets of real world networks.

We also proposed to add dimensionality reduction techniques, namely PCA and ICA, as a preprocessing step to the network comparison. The goal of this step is to use dimensionality reduction as a process to attenuate noise within orbit frequencies, that can be introduced by random number generation (in synthetic networks) or data collection/representation (in real world networks). Results show that this preprocessing improves the performance of NetEmd not only when comparing networks with the same number of nodes and edges, but especially in large datasets containing networks of different sizes.

From an end-user perspective, our extensive testing allows us to recommend guidelines for which version of NetEmd to use in which situation. If the set of networks is homogeneous in number of nodes and average degree, using the largest graphlet size available (size 5 in undirected, size 4 in directed) coupled with ICA (10 to 15 components) leads to the best results. In the cases where there is a large variety in number of nodes or average degree between the networks in the dataset, using smaller graphlets is generally advised. The exception is in directed networks, if the connections within the networks have no reciprocity, then using graphlets of size 4 (with ICA) is more likely to yield a more accurate comparison, due to having a larger set of orbits with non-zero frequency. On the other hand, if the average reciprocity allows using the whole range of graphlets of size 3 and 4, then it is more likely that using graphlets of size 3 coupled with ICA will lead to a more accurate distinction between networks. In this case, 2 components might be enough if the size of the networks are all within one order of magnitude from each other but 5 to 10 components if there is a wider difference.

Chapter 4

Vaccination Discussions in Mumsnet

4.1 Introduction

Social networks are pervasive in modern society, with Facebook, Reddit, Twitter, Instagram and LinkedIn among the most visited sites worldwide. These forms of social media generate immense amounts of data about the interactions of their users, which can be used to inform models of human behaviour. However, the size of these networks generally means it is infeasible to study them in their entirety. For example, it is estimated that over six thousand tweets are sent per second on Twitter. Another consideration is that certain subnetworks—communities within these large networks—may exhibit local behaviour that is different to the global behaviour of the network. As such, it is important when developing models to be able to distinguish between patterns of individual user behaviour and the aggregate statistical properties of such behaviours.

In this chapter, we apply the network comparison tools we developed and presented in the previous chapter to study how people communicate online about vaccination. The World Health Organization named 'Vaccine hesitancy', i.e. the reluctance or refusal to vaccinate despite the availability of vaccines, as one of the 10 major threats to global health in 2019 [187]. Vaccination is a fundamental tool in the prevention of the spread of infectious diseases, and resistance to vaccination has led to outbreaks of controlled diseases like measles [117], putting in danger people who, due to medical conditions, rely on high levels of immunity to avoid contracting these infections. Since vaccine "nonconformers" [17]—parents who do not vaccinate their children or do so with a subset of the recommended vaccines—tend to be proximately located [17], the refusal of vaccination in certain communities reaches values as high as 25% [117].

As might be expected for a health-related topic, vaccination gathers a great deal of

traction in online discussions; it has historically been a controversial topic, and remains a matter of considerable debate [13]. Brunson [17] studied the influence of social and information networks on the vaccination decision and found that the influence of family, friends and medical specialists outweighs that of information sources such as the internet or mass media. However, they also found that "nonconformers" seek these information sources significantly more than their "conformer" counterparts. A study by Davies et al. [29] showed that when searching for keywords "vaccination" and "immunisation", almost half of the results in the most popular search engines are anti-vaccination. This creates a feedback loop for "nonconformers", who seek more information, and thus encounter further information that is anti-vaccination.

The issue of vaccine hesitancy is particularly timely during the COVID-19 pandemic, as the pandemic remains not only a health threat but also a threat to social, economic and general well-being worldwide. Despite evidence of the efficacy of the vaccines developed against the virus [47, 131, 163, 173], concerns rise about the safety of a vaccine due to the hastened process of its development and testing [34], particularly in light of reports regarding rare blood clots linked to the Astra Zeneca vaccine [184]. Although early studies demonstrated that a high percentage of the population was planning to refuse a vaccine against SARS-CoV-2 [128], rollout has been very successful so far, particularly in the United Kingdom (UK), where it has been reported that nearly 50 million second doses and nearly 40 million boosters have been administered, corresponding to uptakes of 86% and 68% [54]. However, research suggests there is still a high level of vaccine hesitancy within specific communities, such as minority ethnic groups [146] and those from most deprived areas [24]. Qualitative studies are still being done to study barriers to the uptake of COVID-19 vaccines [33], but the motivations for hesitancy are often complex and ever changing [137].

The topic of vaccination is, therefore, a source of significant and societal important discussion and debate on social media. The bulk of research on vaccination discussions online is focused on the most popular social media: Facebook [66, 155], Twitter [1, 11, 98, 134] and Reddit [76]. Online chat forums are an often overlooked social medium for vaccination discussions, despite a study by Campbell et al. [20] showing that a significant proportion of participants used discussions forums to find out more about immunisation. The same study found that parents who searched information online were significantly more likely to encounter information that would make them question the decision to vaccinate their offspring, but this likelihood was higher among parents who used discussions forums than those who used Facebook or Twitter. Finally, Rier

[141] highlights the research potential of online support groups embedded internet discussion forums, calling them “natural focus groups" that reflect “segments of public opinion".

The focus of our analysis is on Mumsnet [145], a UK website “by parents for parents". It acts primarily as a website where parents can seek advice on numerous topics, with articles written by other parents, but it also offers a forum functionality that allows users to discuss this same topics. Although there are other websites more focused on vaccination, Mumsnet generates more discussion content regarding vaccination and is more likely to contain a wider range of opinions [165]. The forum content in Mumsnet is also public, anyone can register to comment on any discussion (which appears under a “username", an alias to preserve anonymity) and an account is not required to read or search for past discussions. Prior to July 2019, it contained a subforum dedicated to discussing and asking advice about anything related to vaccination, after its closure the posts within this subforum were merged to the general health subforum. With the COVID-19 pandemic, a subforum dedicated to discussing the pandemic was created, under which a substantial body of discussion about COVID-19 vaccines can be found.

One of the most prevalent techniques to examine vaccinations discussions in social media is *sentiment analysis*, that is identifying and labeling the opinions of social media users towards vaccination and applying quantitative or qualitative methods in order to extract knowledge about those discussions. It is a technique often applied to Twitter data, references [11, 12, 21, 28, 31, 36, 110, 112, 113, 130, 147, 150, 151, 153, 170, 178, 196, 197] are examples but not an exhaustive list. Sentiment analysis has also been applied to Mumsnet before [166].

The majority of approaches using sentiment analysis focus on hand labeling a subset of data (tweets or posts) and building machine learning models to classify the remainder of the data. This is the methodology of Skeppstedt et al. [166] and the author’s major conclusion is that classification in that context is a very hard task, in part due to their small training corpus. This highlights a major issue with the sentiment analysis approach to studying vaccination discussions: obtaining sentiment labeled data is extremely costly, as it is either time consuming or expensive when using platforms such as Amazon Mechanical Turk [19]. There is also a question of reliability of the performance of automatic labeling methods on unseen data, especially when the context and arguments used in the discussion change over time as in the case of COVID-19 vaccine discussions when compared to other vaccines [112]. Finally, using models trained on Twitter data to label non Twitter data is not advisable, as Skeppstedt et al. [166] point

out, Twitter messages are limited to 280 characters so the goal of the message is more easily parsed, which contrasts to the more complex discussions in Mumsnet posts that are harder to interpret.

Given the issues afflicting analysis that rely on sentiment labeling, it is of interest to develop methodologies that can provide as much insight as sentiment or act as a guide to reduce the amount of sentiment labeling required. One such methodology is to consider the interactions between users that are participating in the discussion, creating a network where two users are connected if they interacted with each other according to some criteria. This perspective on vaccination discussions opens up the opportunity to use the wealth of research created in the area of network science to extract knowledge from these discussions. It is even possible to combine the two methodologies, for example, Yuan et al. [197] combine community detection with sentiment analysis. Network analysis has been applied to study the structure of online discussion forums [148, 199], but Mumsnet has not been studied as a complex network before.

By representing vaccination discussions as a network of the actors contributing to the discussion, we propose studying the structure of the network as a proxy for sentiment analysis, leading to the two following research questions:

RQ1: What are the characteristic patterns of interactions between users in vaccinations discussions in Mumsnet and how do these differ between discussions of vaccination in general or when discussing vaccination within the context of COVID-19?

RQ2: Over time, are differences in the network structure correlated to changes in the sentiment about vaccines within Mumsnet?

To identify changes and differences in the structure of the network, we use NetEmd [180] and compare distributions of graphlets. Similarly to *motifs* [109], graphlets can be seen as units of functional behaviour encoded in the network, characterizing distinct patterns of behaviour. In particular, when considering phenomena that spread—such as viruses, ideas and behaviours—the detailed structure of graphlets of the network on which spreading occurs has an important effect on the dynamics of the process [58, 69, 143, 144]. As such, when we observe patterns in the graphlets relating to interactions in online discussions, this will be informative about the manner in which opinions are spread and adopted. By using NetEmd, we can identify changes in the patterns causing differences in the network structure and extract the graphlets responsible for these changes, adding a layer of interpretability that allows the generation of hypothesis to explain the differences in how users are behaving when discussing vaccination online.

4.2 Background

The Mumsnet forum has been the subject of a number of vaccination studies [42, 166, 167] with the measles mumps rubella (MMR) vaccine being a particular focus [165, 181].

Skeppstedt et al. [166, 167] explored 5 threads about general vaccination of children, manually labeling a dataset of 1190 posts as for, against or undecided towards vaccination. Then the authors build a machine learning model to measure the performance of an automatic classifier in this labeled dataset. The machine learning model chosen was a linear support vector machine [27] and the training data was obtained by preprocessing each post into a list of tokens and 2-, 3- and 4-grams. The authors found among the manually annotated posts an even distribution of sentiment, with 38% for and 41% against. The results from the machine learning experiment lead the authors to conclude that their small training corpus makes classification a very hard task.

Ford and Alwan [42] used Mumsnet as a platform for a cross-sectional study of influenza and pertussis vaccination uptake during pregnancy. The study's objective was to learn whether information obtained from online social networks (Facebook, Twitter, etc.) influenced the decision to have these two particular vaccinations. They found that women who use social networks to obtain information were less likely to receive the pertussis vaccination, but this was not the case for the influenza vaccination.

Skea et al. [165] examined two threads about the MMR vaccine from the perspective of parental perception of herd immunity, and the duality between parents not wanting to harm their child versus wishing to avoid harm to others. The authors suggested that vaccine promotional material should "include explanations of herd immunity" as it may influence parents' decisions. Weller and White [181] analyzed posts and threads related to MMR, finding that by 2010 the majority of the forum was in favour of MMR rather than against it.

4.3 Methods

4.3.1 Data Collection

The Mumsnet forum is organized in categories, which are further divided into more specialized topics, called subforums, within those categories. Users create *threads* on these subforums by typing a *thread title* and some text, called the *original post*. Appendix C shows an example of this organization. Other users can reply to the original

post by typing in a text box; these replies, called *posts*, are displayed in chronological order of submission. One of the categories of the forum is *health* and within it, there was a subforum about vaccination, which was closed down in July 2019, and another subforum about the COVID-19 pandemic, created in February 2020. We collect two datasets from Mumsnet, one from each of these subforums.

Prior to the closure of the vaccination subforum, we deployed a web scraper to acquire all threads and posts publicly available in the subforum, spanning back 10 years. This web scraper is simply a *Python* script that given the web page address of a thread, returns all posts within that thread, along with the users, dates and each post's content. We got permission from Mumsnet to acquire this data and the study was approved by the University of Manchester computer science department panel of ethics, with reference 2020-8214-12903. Following the closure of the forum, we set the start date for our data to 1 July 2009, with the last post on this forum dating 28 June 2019. In this time period, there were 1128 threads created, with a total of 31757 posts, written by 5715 users. As a shorthand, this dataset will be referred to as the *vaccine* or *vaccination* dataset.

The COVID-19 subforum contains threads about any topic related to the pandemic, so we filter the threads to scrape according to their title. We include all titles that contain keywords related to vaccination and immunization, as well as COVID-19 vaccine specific terms, for instance, booster or the names of the vaccine manufacturers. There were only 20 threads before 1 November 2020, compared to 81 in November 2020 alone, so we set this as the start date for data analysis. We finalized the data collection on 2 January 2022 and the final dataset only includes posts dated before 2022. In this time period between 1 November 2020 and 31 December 2021, there were 4322 threads created, with a total of 223625 posts, written by 25989 users. Again as a shorthand, we refer to this dataset as the *COVID-19* dataset and every time we refer to the COVID-19 forum, we mean vaccination discussions in particular within this forum.

4.3.2 Data Segmentation by Time

In order to study temporal effects on how discussions are carried out in Mumsnet, we segment the datasets by time, creating new and smaller datasets that span a shorter length of time. One way to accomplish this goal is to create disjoint time intervals so that there is no overlap between the smaller datasets. This method of splitting the datasets gives rise to a dilemma on how to consider which time slice a post belongs to. Consider the follow situation: a thread is created on the 31st of January, half the posts

within that thread are submitted on the 31st and the other half are posted throughout February; assume that the time splicing occurs on the first day of each month. One way to assign data to each time frame is by the date of the thread, but this means that the February dataset does not encompass the discussion that took place in that thread throughout February. Alternatively, half the posts within that thread are assigned to the January dataset and the other half to February. In this case, we are separating discussion that took place within a similar context by imposing an arbitrary boundary; it could be argued that a post from 31st of January is more similar to a post from 1st of February than to a post from 1st of January.

The solution to this issue is to create a rolling window of data, akin to a moving average. This way, we assign posts to data slices according to their post date and ensure that related posts on the edge of the time split are assigned to the same data slice at least once. This method of splicing the data is controlled by two parameters, the *time span*, i.e., how much time each window covers, and the *time jump*, i.e., the amount of time between consecutive windows. The *time overlap* between consecutive windows, i.e. the amount of time in common between consecutive windows, is calculated by subtracting the time jump from the time span.

4.3.3 Network Creation

The way discussion is carried out in a forum is different from other social media platforms like Facebook, Twitter or Reddit in the sense that users interact with threads and messages appear sequentially ordered by time, regardless of relevance to the discussion or user tagging features used. This structure lends itself to the creating of a bipartite network, as two disjoint sets of entities naturally emerge: threads and users, with a user u connected to a thread t if u wrote a post in t. A common technique for analyzing bipartite networks is to create a projection of one of the partitions, for example when one of the modes in the data holds particular interest over the other [14]. The most common way to achieve this projection is assuming that if two nodes in one partition are connected to the same node in the other partition, then they should be connected in the projection. In our case, since we are interested in studying patterns of connections between users, we create a user projection network, where users are connected if they posted in the same thread. Formally, let X be the bipartite adjacency matrix, with users in the rows and threads in the columns, then the user projection network can be

obtained by [14]:

$$a_{ij} = \sum_{t \in \mathcal{T}} x_{it} x_{jt}, \tag{4.1}$$

where $\mathcal{T}$ is the set of threads in the bipartite network and x_{it} represents the number of posts of user i in thread t. In matrix notation, this can be simplified as $\mathbf{A} = \mathbf{X}\mathbf{X}^T$. This transformation produces a weighted undirected network but it is not representative of the reality in forums, as it works under the assumption that all users in a thread read each other's posts. Inspired by Newman [115], we weight Eq. 4.1 to take into account the number of users and number of posts in a thread, similar to collaboration networks where a connection between two researchers is weaker when they publish a paper with a lot of co-authors and stronger when they publish a paper together or with a small number of other co-authors. The strength of the connection is also proportional to the quantity of papers where the researchers are co-authors, with more papers indicating a stronger connection. In our case, two users have a weak connection if they both post in a thread with many posts and a strong one when they participate in multiple threads together. The formula to calculate the projection network then becomes the following:

$$a_{ij} = \sum_{t \in \mathcal{T}} \frac{x_{it} x_{jt}}{(\upsilon_t - 1)\,\phi_t}, \tag{4.2}$$

where υ_t is the number of users in thread t and ϕ_t is the number of posts in thread t. We define a minimum weight for connections between users, below which we assume that the users had insignificant influence over each other. This threshold was chosen as the maximum value that keeps the network connected, i.e., with no isolated nodes.

4.3.4 Network Comparison

The networks we craft from Mumsnet data are a means of representing user posting behaviour and how users interact with each other within Mumsnet. Due to this encoding of information within the network, we expect there to be a relation of equivalence between changes in user behaviour and changes to the structure of the networks. Therefore, analyzing network structure offers us the opportunity to study behaviour within this forum from a different perspective to other approaches. By identifying differences over time in the structure of the Mumsnet networks, differences in the user behaviour can also be made clear.

As explain in detail in Chapter 3, one of the most effective ways to determine

similarity between networks is by comparing them using NetEmd. Upon obtaining the comparison, we can detect differences in the structure of network by looking at the distance between networks from consecutive data slices. If this distance is larger than the median distance between any pair of networks, then we consider that there is a significant difference in the structure of the network. Due to overlapping time frames, differences in the network structure may also take two or three time steps to become apparent.

As graphlets can be seen as small units within a network whose patterns of connections encode real world functionality, we can extract another layer of interpretability from the network comparison results by inspecting which orbits are responsible for the largest difference.

4.3.5 Sentiment Labelling

We previously mention the work of Skeppstedt et al. [166] and their approach to automatic labeling of Mumsnet posts according to their stance on vaccination. The authors highlight the difficulty of this task, particularly in comparison to finding stance in tweets, as posts are usually longer than tweets, given that they are not restricted to 280 characters, so they contain longer and more elaborate discussions. The approach taken by the authors yielded a machine learning model that struggles to learn how to classify such an elaborate corpus, achieving a F_1 score of 0.44.

Starting from Skeppstedt et al.'s approach, we attempted to improve the classification performance of a machine learning in the same labeled dataset, with the intent of applying it to the larger corpus we consider in our work. We found that using a Random Forest [16] improved the classification performance slightly (up to a F_1 score of 0.52). Other approaches like Valance Aware Dictionary and Sentiment Reasoner [71], the Naive Bayes classifier of Salathé and Khandelwal [150] and a FastText model [80] trained on the Crowdbreaks platform [113] all performed worse than adapting the code of Skeppstedt et al. to work with Random Forests.

The poor performance of automatic labeling of sentiment in the training dataset of Mumsnet posts disqualifies this methodology from being applied to the larger corpus of posts, as any analysis done supported on this labeling would have the caveat that the labeling itself is unreliable. Another reason to avoid using a machine learning model trained on Skeppstedt et al.'s data is a *concept drift* [182]. It captures the idea that models trained on a snapshot of time may struggle to generalize for data outside of that time period, particularly in language based models trained on internet data, where

context changes over time. Müller and Salathé [113] showed the existence of such concept drift in vaccination discussions by analyzing a Twitter dataset in the second half of 2018 and at the beginning of the COVID-19 pandemic [112]. The issue of concept drift is particularly relevant when comparing vaccination discussions before and after the COVID-19 pandemic. Changes in the vocabulary, arguments, context and the whole general discourse in these discussions leads to increasing difficulty for a model trained on pre-pandemic data to accurately label a post-pandemic corpus.

The issues we identified with automatic labeling of Mumsnet posts led to the decision to manually label the two datasets we acquired from Mumsnet, to create a ground truth against which we can compare the methodology we develop. It is unfeasible to label each of the >250,000 posts in our dataset; instead, we label threads according to the title and the original post. We classified threads in three labels: *positive*, *neutral* or *negative*. We do not consider negative sentiment to be equivalent to an anti-vaccination position, as it is possible to express negativity towards a vaccine or a vaccination schedule while simultaneously adhering to the policy that recommends taking it. An example of this is starting a discussion about suffering from side effects of a vaccine, without asking for help on how to overcome them, expressing regret over taking a vaccine or doubting the efficacy of a particular vaccine. The following is a non-extensive list of common arguments we marked as negative towards vaccination:

- claims to have refused to be vaccinated and to not plan on being in future;
- wants to vaccinate on a different schedule, delaying recommended or mandatory vaccines;
- shows lack of trust or avoids altogether vaccine cocktails (most commonly the MMR cocktail);
- blames vaccine for illness or expresses regret for taking the vaccine;
- questions the efficacy of the vaccine or the safety of ingredients within the vaccine (excluding allergic reactions);
- supports known opponents of vaccination (e.g., Andrew Wakefield) or anti-vaccination documentaries (e.g., “The Greater Good").

The neutral label captures two types of posts, ones asking for information without expressing a judgment of value towards the vaccine itself or others with conflicting

positive and negative views towards vaccination. Examples of the latter are asking how to deal with side effects or if some side effects are normal, talking about a negative experience surrounding vaccination but not about the vaccine itself or planning to vaccine on schedule but mentioning being afraid or anxious. Finally, we marked threads as positive if they contain arguments like encouraging discussion for more vaccines, details consequences of the disease that can be prevented through vaccination or mentions going out of their way to get a vaccine, for example privately.

The manual labeling process was done by two annotators, who split the annotation load equally, with each annotating half the threads of each dataset. The threads were assigned uniformly at random to each annotator. Prior to annotating, the two annotators labeled 1% of the COVID-19 dataset together (40 threads, chosen uniformly at random), showing high reliability with a Cohen's Kappa score of 0.84 (90% agreement). The annotators also discussed the results to settle discrepancies and agree on the mutual strategy for the remainder of the data.

In order to study the distribution of sentiment within threads of the COVID-19 dataset, we sampled 129 threads (about 3.0% of the number of threads) which contained 6898 posts (about 3.1% of the number of posts), chosen uniformly at random such that they represent at least 1% of the threads and posts in each data slice. The sentiment in all these 6898 posts was labeled using the same criteria used to label threads.

4.3.6 Discordance Index

Under the hypothesis that the proportion of posts of each sentiment within a thread is not sufficient to capture the volume of debate and disagreement, we propose an alternative metric to measure this level of disagreement, which we call the *discordance index*. The main idea behind this metric is to measure disagreement on a local level by measuring how often sentiment changes in small groups of consecutive posts and use the number of sentiment changes as a proxy for the level of discordance, weighing changes from positive to negative (or vice-versa) more than changes involving neutral sentiment.

An example that motivates this metric is the following: consider a thread with N posts, where the first $N/2$ posts are of negative sentiment and the last $N/2$ are positive. By considering only the proportion of positive to negative posts, this thread would appear to contain a large amount of disagreement between users as half the posts express a diametrically opposed view of each other. However, by inspecting how the sentiment

is distributed over the thread, we actually find that most users are agreeing with each other, with the exception of the shifting point from negative to positive. On the other hand, if a thread with N posts has $N/2$ negative posts interleaved with $N/2$ positive posts, then it is much more likely that users are constantly disagreeing with each other throughout the thread. In both cases the proportion of negative to positive posts is the same, but the way discussion is being carried out is very different.

Our proposed metric *discordance index* attempts to capture this distinction of local variety of sentiment within a single thread. The way it works is by considering all sets of D consecutive posts, such that in a thread with N posts there are $N-D+1$ sets of D posts. For each set, we compare every unique pair of posts within that set (leading to $D*(D-1)/2$ comparisons), if one post is positive and the other is negative, we assign a distance of 2; if a post is neutral and the other is positive or negative, we assign a distance of 1; finally, if the posts are of the same sentiment, we assign a distance of 0. The discordance within a set of posts is the sum of distances between each pair of posts; the discordance of a single thread is the sum of discordances of all sets, divided by the maximum discordance for that set size. This maximum discordance of a set of posts can be calculated by generating all the possible combinations of sentiment for a fixed set size and comparing them to find the combination with highest discordance. Finally, the *discordance index* of a thread is the average discordance across multiple window lengths, which we vary between 2 and 5.

4.4 Results

4.4.1 Data Granularity

We start by identifying appropriate window overlaps to split each dataset over time. Although answering the research questions we posed in the introduction to this chapter should not depend on these parameters, they are responsible for controlling how much change and variability are in each data slice. For instance, picking a window span too long may lead to a uniform sentiment throughout the whole dataset, whereas a window span too small can lead to too much variability between consecutive time windows such that any difference becomes meaningless.

The first result we use to support the decision is looking at how long each user spends on the forum (Figure 4.1) and how long each thread remains active (Figure 4.2). These figures show that users behave differently in the two subforums. On the vaccine

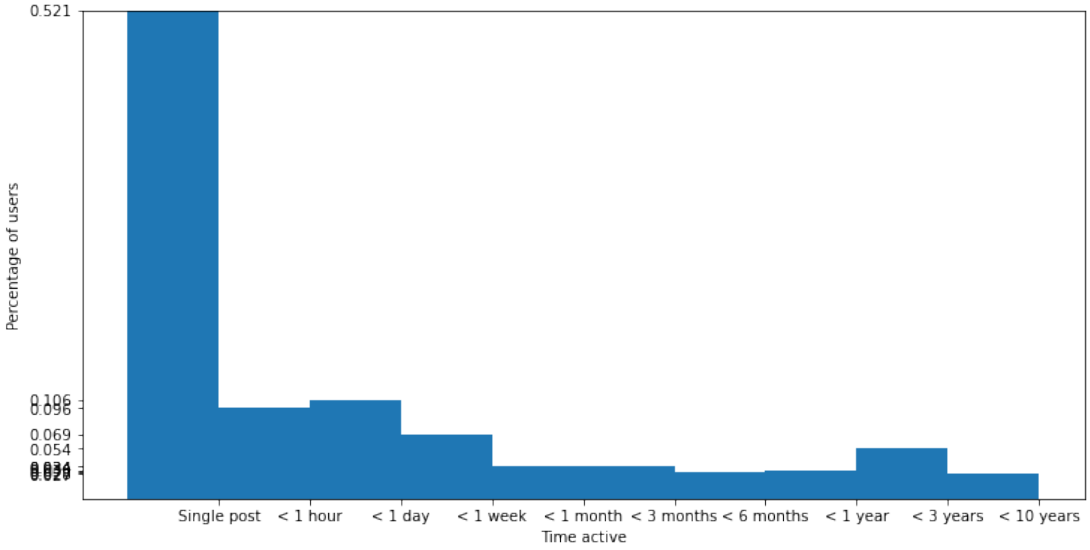
dataset, we find that 52% of users post only once and close to 10% of threads get no replies. This contrasts with the COVID-19 dataset, where 37% of users only have one post and 4% of threads did not get a single reply.

Looking at user active times, results show that users engage with the forum for a shorter amount of time in the vaccine forum compared to the COVID-19 forum, 80% of users in the vaccine forum are active for less than a week. On the other hand, in the COVID-19 forum, if we exclude users with a single post, then it is more likely that a user will be active longer than a week (42% of users) than shorter (20% of users). It is expected that the COVID-19 forum shows more engagement as the pandemic was an extremely disruptive event worldwide, whereas the issues brought up in the vaccine subforum are more likely to be short lived. Despite this, results highlight that there is a core group of users in the vaccine forum with long term engagement, 8% of users were active on this forum for longer than a year and close to 3% were active for more than 3 years.

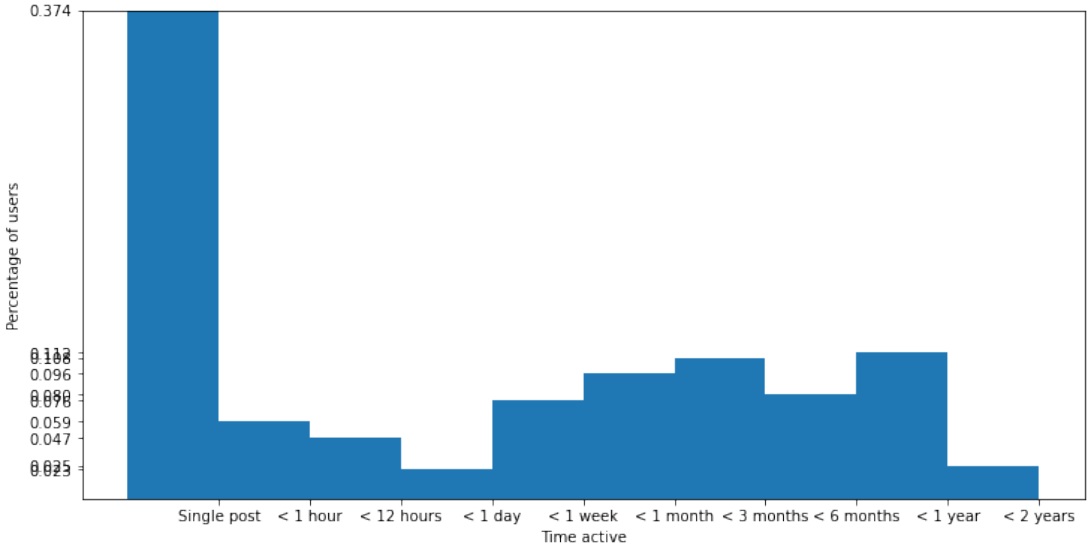
Inspecting the thread active times in the COVID-19 forum highlights how fast information changes during a global pandemic, particularly in the modern age where information is easily disseminated and accessible. Results show that 42% of threads last less than 12 hours (46% if taking into account posts without replies); close to 28% of threads last less than a week but longer than a day and only 11% of threads last longer than a week. This is likely because as new information about the vaccines was being released, older discussions lost their relevance quickly in favour of new facts. This contrasts with how discussions were being held in the vaccine forum, with only 62% of threads lasting less than a week, with 1.1% of discussions even lasting longer than 3 years, compared to only 1.3% of COVID-19 discussions lasting longer than 3 months.

These two sets of figures lead us to conclude that we should favour a longer time span of each data slice from the vaccine dataset, because many users post for a short amount of time but their influence can be long lasting due to a substantial amount of discussions lasting multiple months. On the other hand, the time span of data slices from the COVID-19 dataset should be smaller, as threads tend to last a small amount of time and users showing signs of repeated engagement with the forum.

Based on these observations, we narrow down the candidates for window span and window jump for each dataset and show how the number of users, threads and posts evolves over time under each combination of span and jump in Figures 4.3 and 4.4, for the vaccine and COVID-19 dataset, respectively.

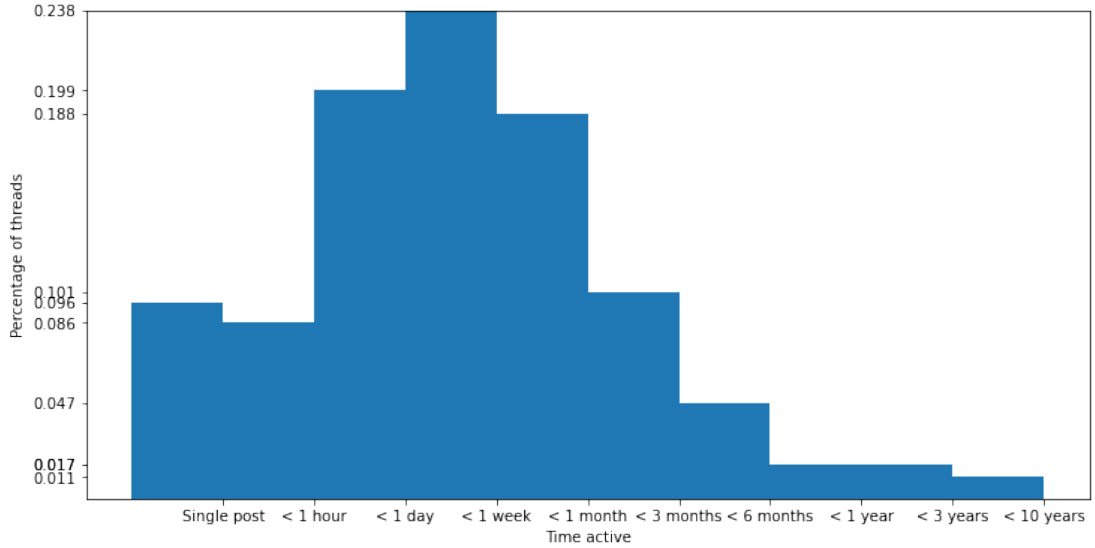


(a) Vaccination subforum

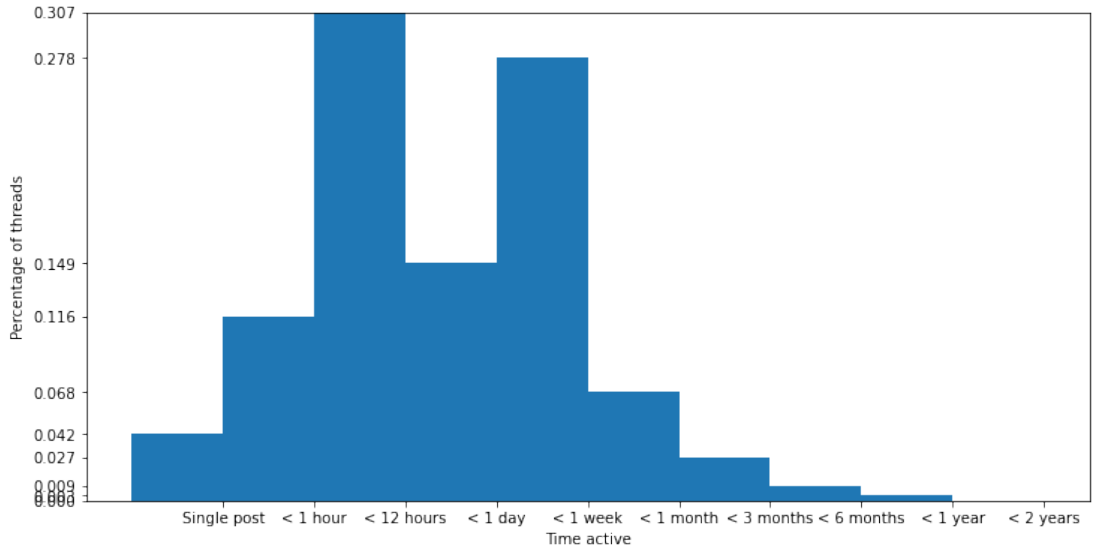


(b) COVID-19 subforum

Figure 4.1: User active times: distribution of how long each user remains active in each forum, obtained by taking the time difference between the date of their first and last posts.



(a) Vaccination subforum



(b) COVID-19 subforum

Figure 4.2: Thread active times: distribution of how long each thread remains active in each forum, obtained by taking the time difference between the date of their first and last posts.

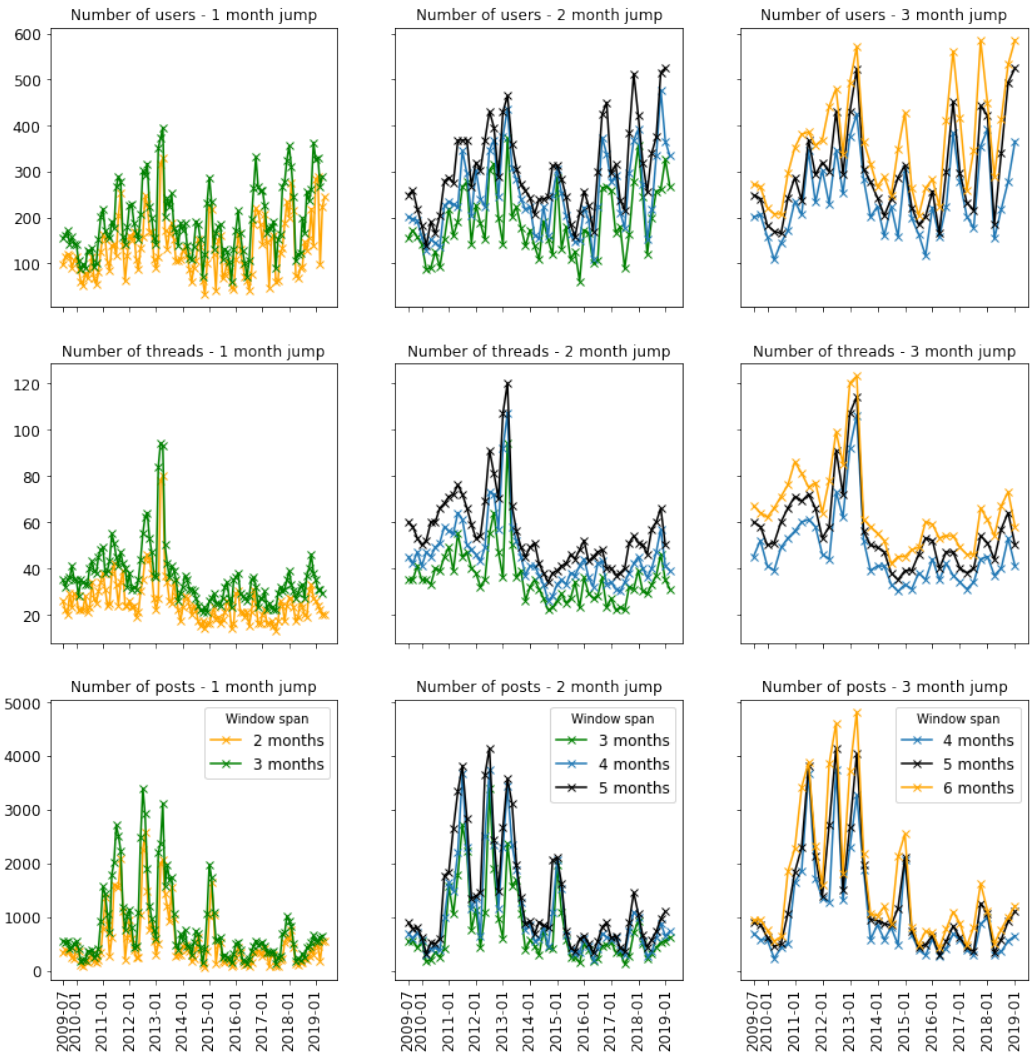


Figure 4.3: Number of users, threads and posts for different combinations of time jump and time span to define the time granularity in the vaccine dataset.

Inspection of Figure 4.3 highlights the similarities between each combination of parameters to split the vaccine dataset. We find the most appropriate window jump value to be 2 months, achieving a balance between sufficiently small granularity that minimizes extreme jumps in the values of these statistics, while enough difference can be observed between consecutive time slices. The vaccine forum does not have enough traffic that a 1 month or smaller jump leads to enough variability between consecutive data slices; this is supported by noticing that peaks in each metric all contain several data points in close proximity, so they could be interpreted as repeated data that makes identifying differences over time harder. On the other hand, using a 3 month jump and particularly with longer window spans, leads to the omission of more subtle differences and hinders pinpointing with accuracy when changes has occurred, along with events that led to the change.

Fixing the window jump narrows down the possibilities for the values of window span. A span of two months leads to no overlap, so we can focus instead on values longer than this. A span greater than 5 months leads to overlaps at least twice as big as the time jump, which can make identifying differences between consecutive data slices harder. Between 3, 4 and 5 month spans we do not observe significant dissimilarity between the shapes of the curves for each statistic. We pick a 4 month span to ensure that each month is equally represented in the dataset, which only happens if the time span is a multiple of the time jump. For example, if the span is 3 months, odd numbered months are represented in two data slices but even numbered months are only represented in one.

For the COVID-19 dataset (Figure 4.4), we want the data granularity to reflect the fast pace of new information that users are exposed to, so small window overlaps are desired. Recalling our previous remark that short window spans are a better fit considering the distribution of thread active times, we can rule out window spans of 2 months or over; with these time spans, each data slice would represent over 10% of the dataset. By desiring small window overlaps, we can also discard the 1 month span combination with 1 week jump, as each consecutive data slice would share 75% of the data. Finally, between the 1 month span combination with 2 weeks jump and the 2 weeks span combination with 1 week jump, we choose the former as it is closer to the timescale of policy changes that may influence discussions. The monthly span also makes it easier to reason about the time frame that each window spans, as we find it to be more interpretable to separate results according to the month they take place in.

To summarize, for the vaccine dataset, we pick a 4 month span with 2 month time

Figure 4.4: Number of users, threads and posts for different combinations of time jump and time span to define the time granularity in the COVID-19 dataset.

		Vaccine	COVID-19
# Networks		59	27
Number of Nodes	Min.	101	1873
	Median	222	3855
	Max	469	5743
Number of Edges	Min.	264	25000
	Median	928	94750
	Max	4616	219774
Density	Min.	0.015	9.14×10^{-3}
	Median	0.036	0.014
	Max	0.052	0.027
Average Degree	Min.	4.6	25.7
	Median	8.1	50.3
	Max	19.7	84.8
Clustering Coefficient	Min.	0.40	0.71
	Median	0.70	0.78
	Max	0.93	0.83

Table 4.1: Summary statistics for the datasets of networks from the vaccination and COVID-19 datasets.

jump, leading to a 2 month overlap between consecutive data slices. For the COVID-19 dataset, we pick a 1 month span with 2 weeks time jump, leading to a 2 week overlap between consecutive data slices.

4.4.2 Temporal Network Comparison

Upon splitting the datasets and creating a network by the process described in Section 4.3.3 for each time window, we obtain two datasets of networks, whose summary statistics are shown in Table 4.1.

The networks in each dataset correspond to users connected through their interactions in a medium that does not change and created using the same the methodology, therefore it is reasonable to assume that all the networks have the same generation mechanism. Then, this means that a *network comparison* method like NetEmd, that can separate networks according to their generation mechanism and is made for identifying common organizational principles, should output that the networks within each dataset are similar to each other. If the networks are dissimilar according to the network comparison method, then the way that users are interacting within the forum is also different.

Considering the disparity in number of nodes and average degree over time in both

datasets, results from Chapter 3 indicate that the most suitable version of NetEmd is *PCA_NetEmd* with size 4 graphlets and 90% explained variance, but we find a high level of agreement between this parametrization and the original NetEmd formulation. The symmetric distance matrix returned by NetEmd containing the distances between each pair of networks can be visualized through a heat map. This visualization technique helps highlighting clusters of networks that are similar and quickly identifying networks that differ substantially from their predecessors and successors. The heat map for the vaccine dataset is shown in Figure 4.5 and for the COVID-19 dataset in Figure 4.6. On the axis of each heat map, we include the *start date* of the time window each network is built from. For example, on the vaccine dataset, the first network, labeled as “07-2009", has data spanning from 1 July 2009 until 31 October 2009; on the COVID-19 dataset, the first network, labeled as “11-2020", has data from 1 November 2020 until 30 November 2020, the second network (not labeled) from 15 November 2020 until 15 December 2020, the third (labeled “12-2020") from 1 December 2020 to 31 December 2020 and so on.

4.4.2.1 Vaccination forum

Inspection of Figure 4.5 reveals two broad groups of networks, before and after May 2015. Prior to this date, the heat map shows two networks that are substantially different compared to neighbouring time slices, one in March 2012 and the other in September 2014; but the remaining networks before May 2015 possess a high degree of similarity among them, as indicated by the darker hue. After May 2015, the network structure changes more frequently, as can be noted by the lighter hue colouring the distances between these networks. Although they are nonetheless more similar to each than to networks before May 2015, the distances among elements of this post May 2015 cluster are, on average, higher than between the networks that compose the pre-May 2015 cluster. Also note that the transition between these clusters is gradual, starting with the network from May 2014 until the network from March 2015, these intermediate networks are similar to elements in both clusters.

Some of these differences in network structure occur simultaneously with controversies surrounding vaccines. By inspecting post content in the data slice where we detect network differences, we found evidence that these news stories and controversies penetrate discussions in the forum. For instance, this is the case with the differences observed in the network starting in May 2014, that includes a study by Hooker

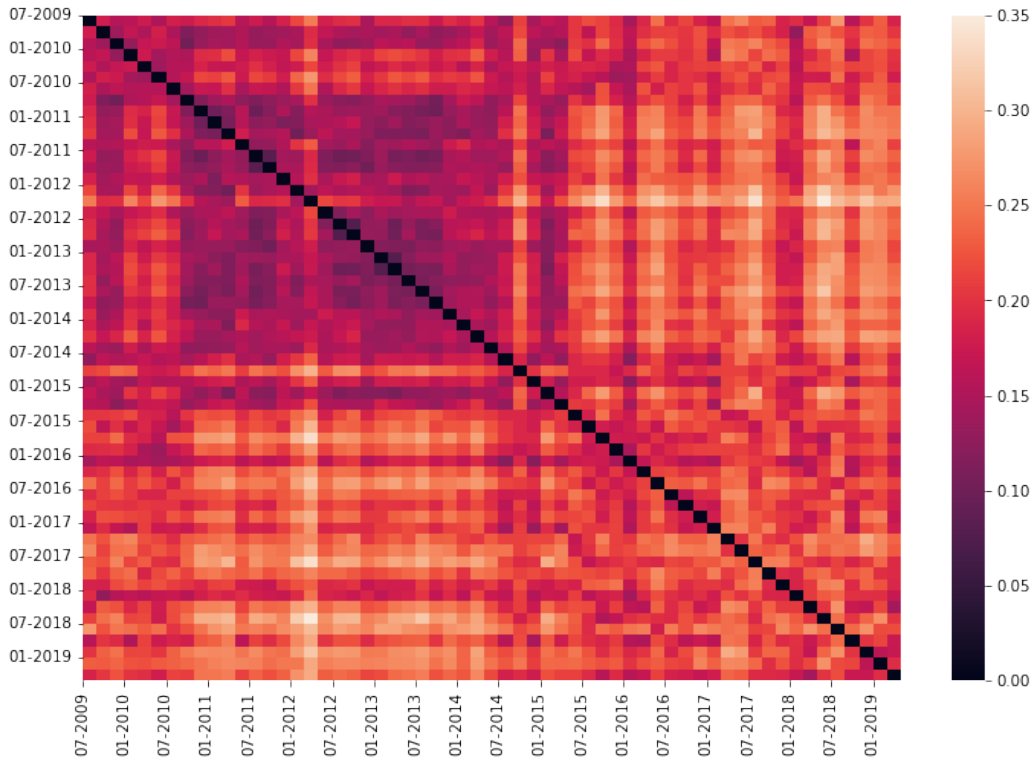


Figure 4.5: Heat map of network distances between each pair of networks in the Mumsnet vaccine dataset. Distances are calculated using *PCA_NetEmd* with size 4 graphlets and 90% explained variance.

[67] where the connection between the MMR vaccine and autism is investigated. Another example is the network starting in May 2015, during which a child whose parents opposed the vaccinations died of diphtheria in Spain. We show examples of threads referring to these events in Appendix C. Other differences may not be related to world events regarding vaccination, but instead with controversial topics started by users in Mumsnet that generate a high volume of discussion, for example the difference between the networks of March 2012 and May 2012 could be related to a discussion regarding the safety of aluminium in vaccines.

To quantify the changes we observe in the heatmap, we compare the distance value between networks of consecutive time windows with the median of the values in the upper triangle of the distance matrix. We also compare this median against the distance value between networks from time windows that are 2 jumps apart, pairs of networks with no overlapping data but spanning a continuous time frame. Our measure for significance of difference between consecutive networks is if their distance is greater than the median of distances between all pairs of networks. Prior to May 2015, we only find significant differences in the same time windows we exposed before: March 2012 to May 2012 and May 2014 to September 2014. After May 2015, we detect 8 significant differences between networks from consecutive time slices and 11 when comparing with 2 jumps apart.

The NetEmd measure between two networks is calculated using the average distance for each orbit, therefore it allows us to pinpoint which graphlets are driving changes in the network structure by inspecting the distances between distributions of each orbit. The heat maps of network distances for each individual orbit are shown in Figure D.2. We focus on two time periods as an example:

- The difference between the networks starting in March and May 2012 is primarily connected with the distributions of orbits 1, 6 and 9 (refer to Figure 2.1). The common element between these 3 orbits is that within their graphlet, they are connected to a node with no other connections.

- The distinction between the clusters before and after May 2015 is largely connected with differences in orbits 0 (the degree distribution), 3 and 14, which are all orbits part of cliques.

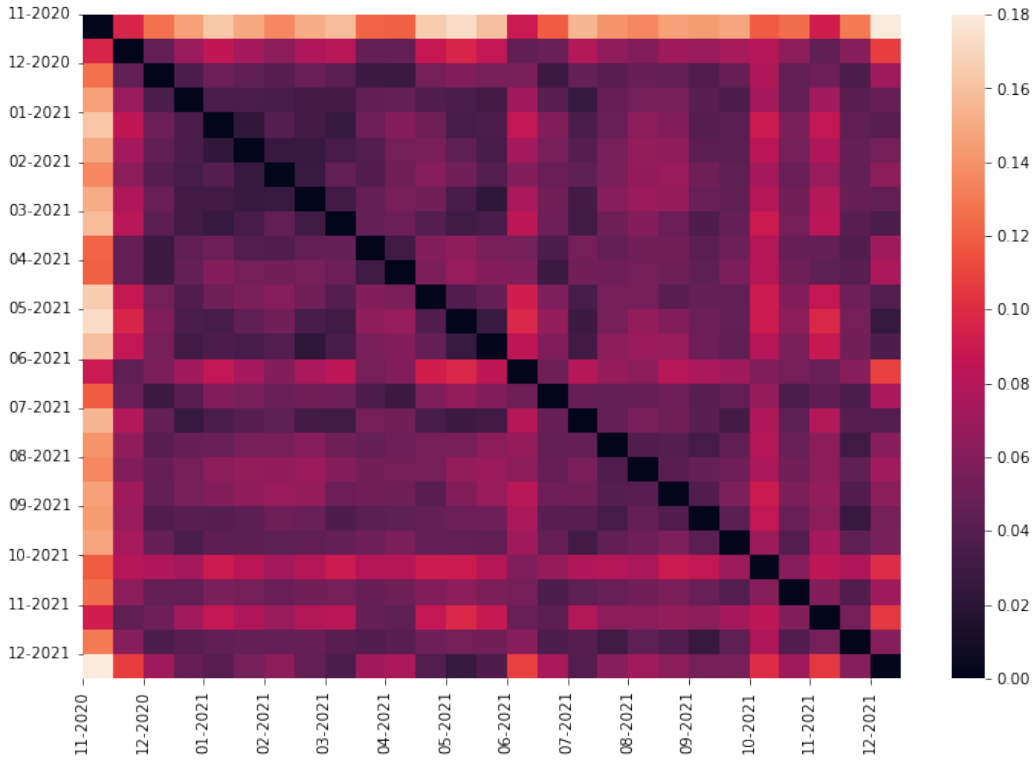


Figure 4.6: Heat map of network distances between each pair of networks in the Mumsnet COVID-19 dataset. Distances are calculated using *PCA_NetEmd* with size 4 graphlets and 90% explained variance.

4.4.2.2 COVID-19 forum

Comparing networks from the COVID-19 dataset yields an altogether different pattern of similarities, as shown in Figure 4.6. The heat map of distances is dominated by darker hues, meaning that the majority of networks are very similar. The most flagrant exception to this high level of similarity is the first network, which spans the month of November 2020 and is the only network in this dataset that precedes the announcement of the first COVID-19 vaccine being approved. Recall that the first vaccine to be approved in the UK was the Pfizer vaccine on 3 December 2020, so the second network in this dataset, starting on 15 November 2020, already includes this period. This second network also shows a high degree of dissimilarity when compared with the remaining networks in this dataset, but not to the same extent of the first one.

Two other periods of time stand out when analyzing the heat map, the networks starting in 1 June 2021 and 1 October 2021. The network starting in 1 June 2021 coincides with the Delta variant becoming the dominant strain of COVID-19 in the UK, which was thought to be more resistant to COVID-19 vaccines than previous

variants, in particular when only a single dose had been administered, at a time when approximately 50% of the population had received the second dose. The network with start date of October 2021 coincides with the beginning of the booster roll out in the UK, which seems to be the leading topic of discussion during this time as “booster" was the sixth most common word among all posts during October 2021 (when removing stop words, after “COVID", “vaccine", “people", “think" and “vaccinated").

Applying the same analysis as before–comparing the distance between networks from consecutive and 2 jumps apart time windows against the median distance between all pairs of networks–reveals new points in time where there is a significant difference in network structure. In addition to the previous three time windows, we also find that the networks with start date 15 March 2021 and 1 April 2021 to be significantly different from their neighbouring time periods. These two networks coincide with the blood clots controversy surrounding the Astra Zeneca vaccine. Finally, we also find a significant difference between the network starting in 1 November 2021 and the network starting 1 December 2021. This difference is likely related to the emergence of the omicron variant and the government’s push for more boosters to increase protection against the new variant. Examples of threads about these issues can be found in Figure C.5 in Appendix C.

When inspecting the differences between distributions of each orbit, visualized in Figure D.3, we find that the network of 1 November 2020 is significantly different from the other networks according to every orbit except orbit 8, according to which this network is similar to the following two networks (15 November and 1 December 2020). Similarly, the difference between networks of 15 May and 1 June 2021 is significant according to all orbits. On the other hand, we find that the difference in the network starting on 1 October 2021 is driven primarily by changes in the distributions of cliques, through orbits 3 and 14, but we also observe smaller changes in orbits 0, 10 and 13.

We conclude this section with a final remark about the differences between the vaccine and COVID-19 datasets, a comparison that we show in Figure D.1. We find that there is a clear distinction between the two sets of networks in the way they are organized, which indicates distinct user behaviours in each forum. This distinction is more tenuous when comparing the first network from the COVID-19 dataset, that shows a comparable value of dissimilarity with the networks from the vaccine dataset as with the other networks from the COVID-19 dataset. This first network has an average distance to other COVID-19 networks of 0.18 and to vaccine networks of

0.21; the other COVID-19 networks have an average distance of 0.06 to COVID-19 networks and 0.20 to vaccine networks.

By analyzing the differences between the vaccine and COVID-19 datasets for each individual orbit (Figure D.4), we find that in 6 of the 14 orbits the COVID-19 dataset is more similar to the cluster of similar networks in the vaccine dataset before May 2015 than this cluster is similar to the networks of the vaccine dataset after May 2015. However, the opposite only happens in orbit 9. This highlights how fundamentally different is the way that users behave before and after May 2015 in the vaccination forum, to the point where networks whose size differs by more than an order of magnitude are organize in a more similar way than networks of comparable size.

4.4.3 Sentiment Analysis

We turn our attention to exploring the sentiment of Mumsnet in regards to vaccination. Our sentiment labeling methodology complicates defining what is the sentiment of a social network like Mumsnet, as we only have access to the sentiment that seeded each individual discussion thread. Therefore, we combine three metrics to summarize sentiment in each time window:

- Number of threads: the most direct metric from our labeling process, we measure the proportion of threads labeled with each sentiment.

- Number of posts in thread: compare how many posts are there in threads of each label. It allows us to gauge how popular each type of thread is, to identify cases where, for example, there are many threads of a sentiment label but each thread has very few posts; or the opposite, where there are only a few threads of a sentiment label but each thread receives a disproportionate number of replies.

- Number of users: we measure which types of threads attract more users and how often users post in threads of different sentiment. This metric is useful to understand how much information is flowing between threads of different sentiments.

For each of these metrics, we measure how the proportion of positive to negative, negative to positive and neutral to non-neutral changes over time.

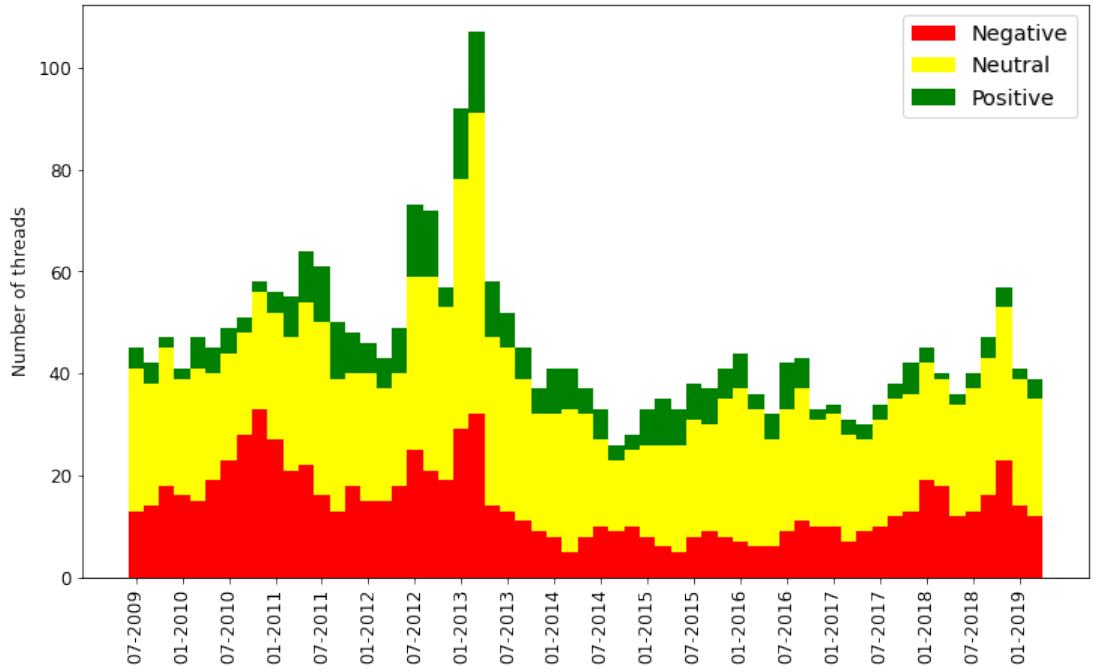
4.4.3.1 Vaccination forum

Starting by the vaccination subforum, we show how the number of threads of each sentiment (Figure 4.7), the number posts in each type of thread (Figure 4.8) and how many users post in threads across each combination of thread type (Figure 4.9) changes over time.

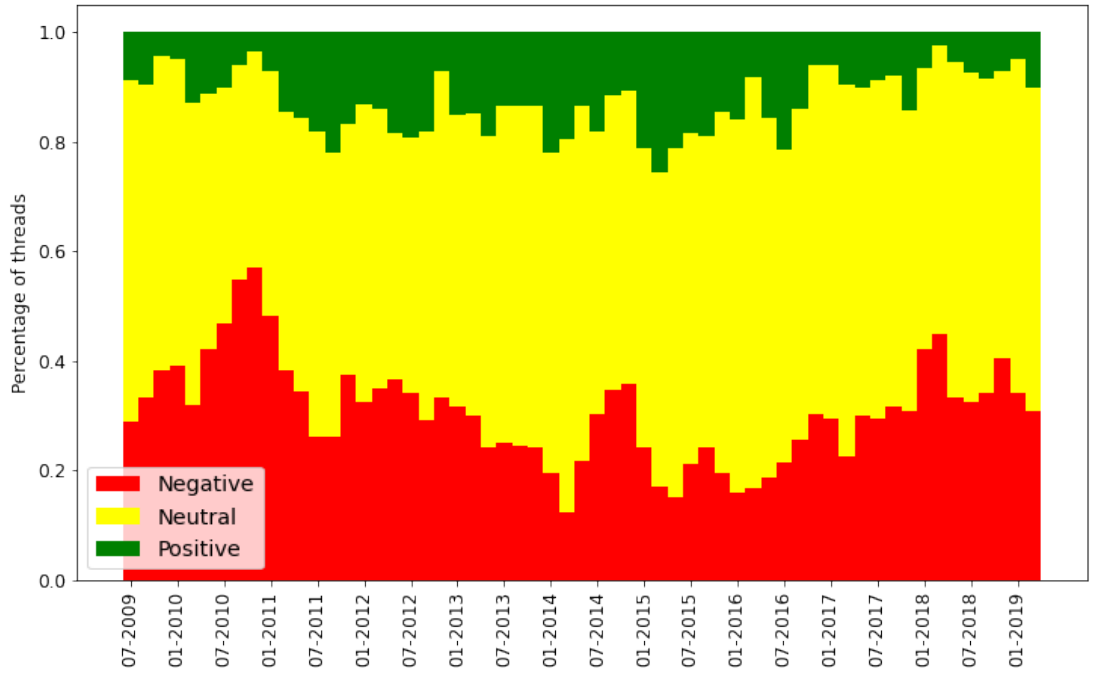
We find that the most common sentiment when opening a new discussion thread is neutral, with 58% of threads labeled this way. This is expected because one of the primary uses of the Mumsnet forum is for parents to seek advice from other parents, therefore many of these threads we marked as neutral are questions that do not incite much discussion. Evidence that supports this is that neutral threads tend to have shorter active times, with 70% lasting less than a week, against 56 and 52% for positive and negative, respectively. The second most common sentiment is negative, with 29% of discussions starting from a negative point of view towards vaccination. The proportion of negative threads is particularly high before March 2013 and after November 2017, respectively with 37% and 36% of negative threads during these time periods. The least common type of thread is positive with 13% of all threads, which is not surprising as those who are content with the status quo (high rates of vaccination) do not feel compelled to support this status quo as much as those who have problems to expose those problems.

Inspecting the number of posts in threads of each sentiment reveals that the majority of posts were written as replies to negative threads, with 44% against 36% in neutral threads and 20% in positive threads. In spite of this, positive and negative threads receive on average a similar amount of replies, 44 and 42 respectively, much higher than the average 17 replies per neutral thread. The visualization of Figure 4.8 highlights how there are time periods dominated by discussions of each sentiment, for example from September 2017 to March 2018 78% of posts were replies in negative threads but from May 2015 to May 2016 this proportion decreases to 28%.

Figure 4.9 shows that the majority of users post only in one type of thread, 89% of users post only in negative or positive or neutral threads. The biggest crossover between types of threads is neutral and negative, accruing an average of 6% users over time, against 1% that post in both positive and negative, 2% in positive and neutral and 2% in threads of all sentiments. We find that after July 2015, the proportion of users that post in threads of all sentiments or in positive and negative threads decreases significantly; prior to this date these percentages are 3.5% and 1.8% respectively, decreasing to 0.6% and 0.5% after July 2015.

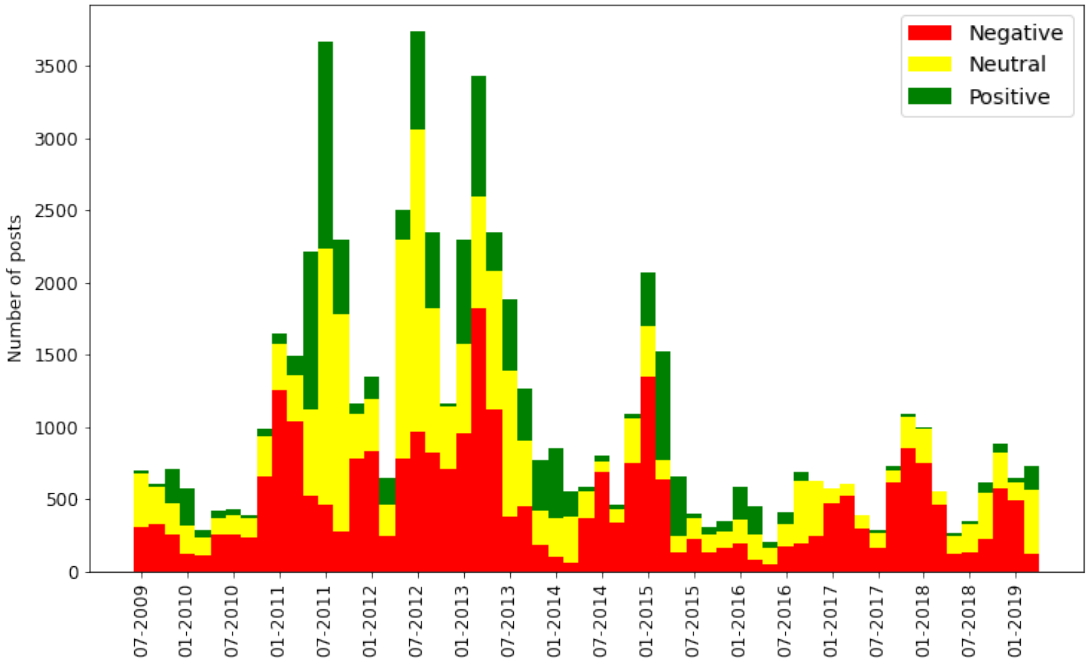


(a) Number of threads.

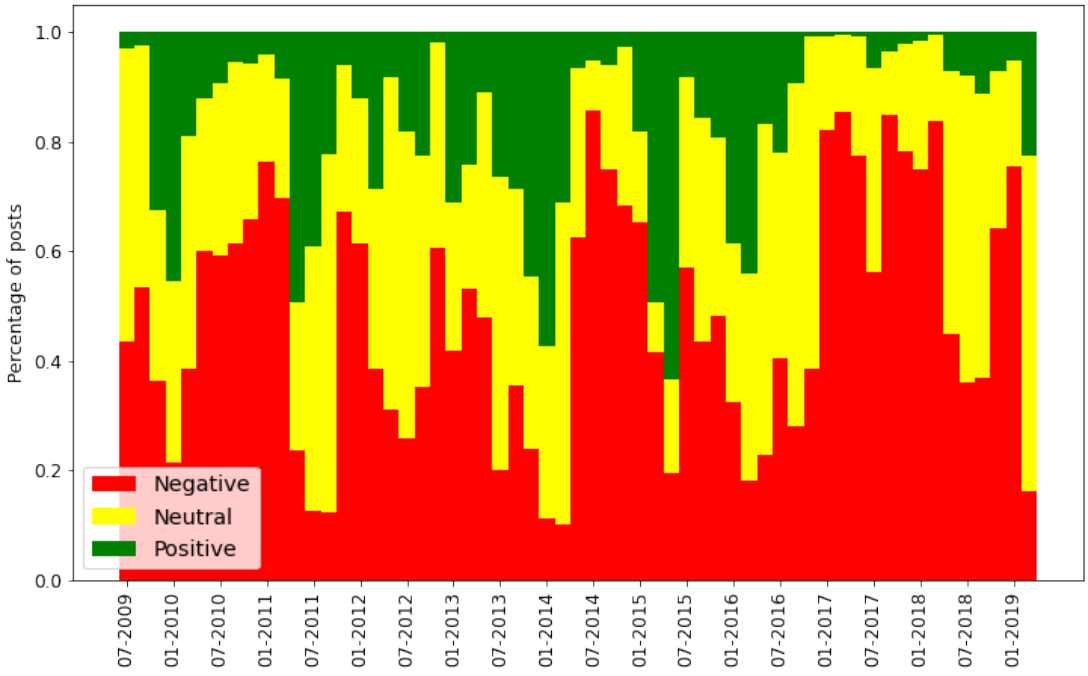


(b) Proportion of threads.

Figure 4.7: Number and proportion of threads of each sentiment label, for each data slice in the vaccination dataset.

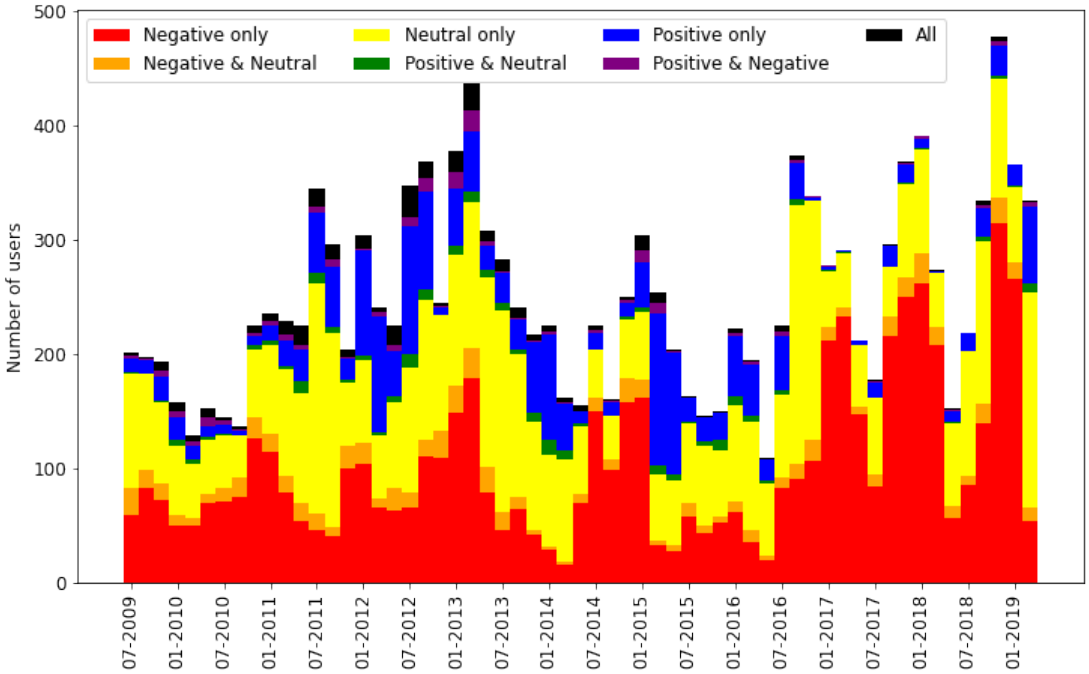


(a) Number of posts.

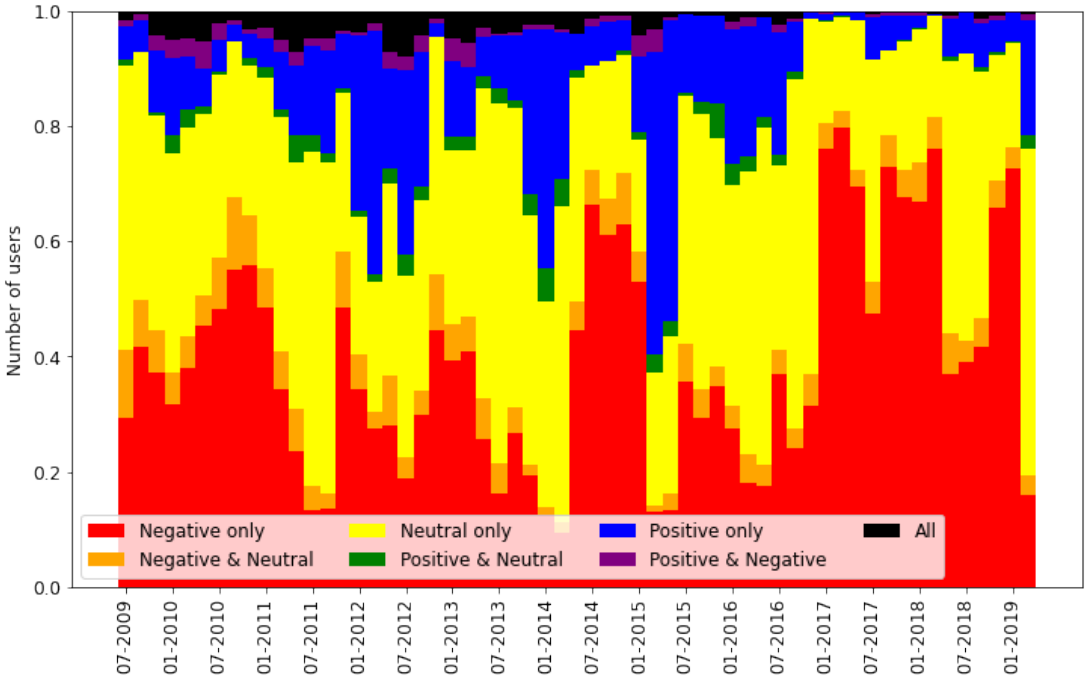


(b) Proportion of posts.

Figure 4.8: Number and proportion of posts in threads of each sentiment label, for each data slice in the vaccination dataset.



(a) Number of users.



(b) Proportion of users.

Figure 4.9: Number and proportion of users in threads of all combinations of sentiment label, for each data slice in the vaccination dataset.

These numbers are biased by the majority of users who post in only one thread. Across all data slices, on average only 15% of users post in multiple threads but these users are responsible for 47% of posts. When we consider only these users that post on multiple threads, the majority of them (69%) post in threads of different sentiments, with 1% posting only in positive threads, 19% in negative and 11% in neutral. As before, the most common combination is posting in negative and neutral threads, with 38% of users doing so, followed by posting in threads of all sentiments, with 14% of users.

To summarize, even though Figure 4.9 shows that a small proportion of users post in threads of different sentiments, when restricting the set of users to those that post in more than one thread, they are more likely to post in thread of different sentiments. As these users are responsible for nearly half of the posts on the board, it is reasonable to assume that there is information flowing between threads of different sentiments.

We are interested in knowing if the shifts we observe in sentiment correlate with differences in network structure. Thus, we need a systematic and quantifiable way to detect that sentiment in a time window is significantly different from other time windows. To achieve this, for each of the metrics above, we calculate the proportion of positive to negative, negative to positive and neutral to non-neutral and Z-score (subtract the mean and divide by the standard deviation) this proportion. When the Z-scored proportion takes values above 1 or below -1, we consider the respective time window to contain a significant change in sentiment. Figures D.5, D.6 and D.7 show this Z-score over time for threads, posts and users, respectively. These figures also contain all the network differences we mention in the previous section, marked as a black dotted line on index i when the change occurs from index i to $i+1$ and a gray dotted line when the change occurs from index i to $i+2$. In the cases when both black and gray dotted lines ought to be present, we give precedence to the difference between consecutive time windows and show only the black dotted line.

These figures show a total of 40 time windows that fit such definition of significant change in sentiment, out of a total of 59. We narrow this number down by considering only time slices that are marked as significant according to different metrics related to the same sentiment, for example an increase in the proportion of neutral threads and posts in neutral threads. This narrows down to 20 time windows with significant change in sentiment, compared against 16 differences in network structure.

We separate the differences in network structure in 6 groups:

- The difference between the network starting in March 2012 and the network

starting in May 2012 is correlated with a decrease in the proportion of neutral threads (Figure D.5c) and an increase in the proportion of posts in neutral threads (Figure D.6c). There is also a significant increase in the proportion of users that post in multiple threads of different sentiments during this time (Figure D.7d).

- The differences between the networks of May 2014 to September 2014 and of September 2014 to January 2015 do not seem to match with significant differences in sentiment during this period.

- The changes in network structure in March 2015 and May 2015 can be associated to increases in the proportion of positive threads (Figure D.5a), posts in positive threads (Figure D.6a) and users posting only in positive threads (Figure D.7a).

- The differences in March and May 2016 are correlated with increases in the proportion of neutral threads (Figure D.5c), posts in neutral threads (Figure D.6c) and users posting only in in neutral threads (Figure D.7c).

- Between November 2016 and July 2017, we record differences of network structure in 5 consecutive time windows and during this time we observe an increase in the proportion of neutral threads (Figure D.5c), an increase in the proportion of posts in negative threads (Figure D.6b), an increase in the proportion of users posting only in negative threads (Figure D.7b) and a decrease in the proportion of users posting in multiple threads of different sentiments (Figure D.7d). There is also a significant shift from the time slice starting in November 2016 to January 2017 in the number of posts and users in neutral threads (Figures D.6c and D.7c), from a proportion significantly higher than the average to significantly lower.

- The group of network differences starting on January 2018 until July 2018 is related to an increase in the proportion of negative threads (Figure D.5b), posts in negative threads (Figure D.6b) and users posting only in in negative threads (Figure D.7b).

We also find periods of time where there is change in the sentiment that is not reflected in differences to the network structure. In particular, we observe a high proportion of positive threads (Figure D.5a), posts in positive threads (Figure D.6a) and users posting only in positive threads (Figure D.7a) in January and March 2014. The final time period that fits our definition of sentiment change but does not align with differences in network structure is July and September 2011, where we detect an higher

representation than usual of posts in positive threads (Figure D.6a), posts in neutral threads (Figure D.6c) and users in neutral threads (Figure D.7c).

4.4.3.2 COVID-19 forum

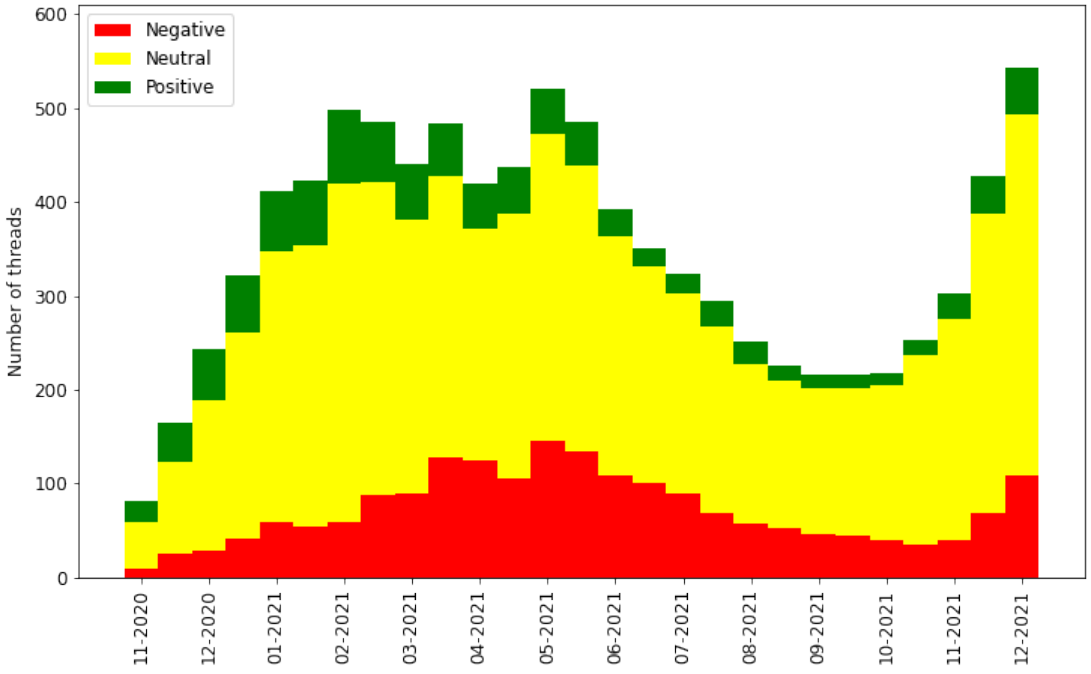
Proceeding to the COVID-19 subforum, we show how the number of threads of each sentiment (Figure 4.10), the number posts in each type of thread (Figure 4.11) and how many users post in threads across each combination of thread type (Figure 4.12) changes over time.

Similarly to the vaccination forum, the most common sentiment when starting a thread about vaccination during COVID-19 is neutral with 69% of threads. A major difference to the vaccination forum is that threads with neutral sentiment are the most common in all data slices, whereas in the vaccination dataset there were time windows with more negative threads than neutral. The reason for this prevalence of neutrality is that many threads about vaccination during the pandemic are not discussing the actual vaccine but instead issues surrounding the vaccination campaign, such as logistics or rules for vaccinated people, so the original post of these threads does not contain any judgement of value towards the vaccine, which we marked with neutral sentiment.

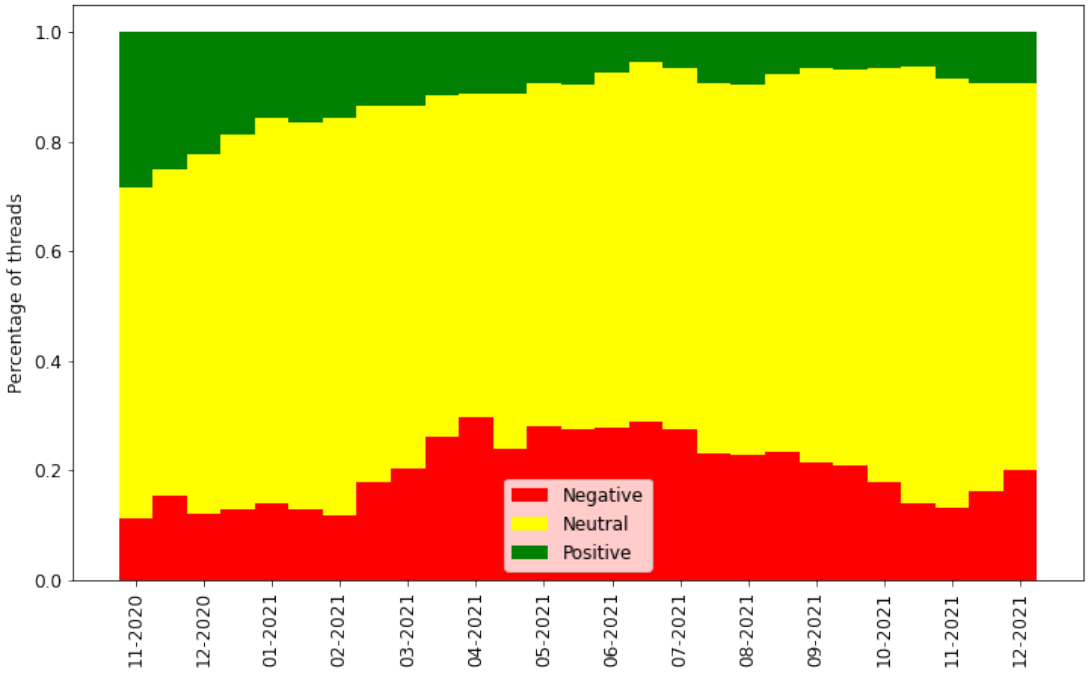
As before, the second most common sentiment is negative with 20% of threads. The negative sentiment is particularly high between March and July 2021, mainly due to the Astra Zeneca blood clot issue and users complaining about side effects of vaccines. Finally, the least common sentiment is positive with 11% of threads. However, the first four time windows whose start date takes place in 2020 have a much higher proportion of positive threads compared to the remainder of the dataset, with an average of 23% of threads over this time period.

In contrast to the small number of positive threads, we find that positive threads attract the highest number of replies, averaging 70 replies per time window, against 60 in negative threads and 39 in neutral. Despite attracting more replies, the low amount of positive threads means that posts in positive threads are nonetheless the minority of all posts, averaging 17% of all posts over time, as pictured in Figure 4.11. The popularity of individual positive threads can be explained by “positivity" threads, created to be a safe space where users can feel good about the progress of the vaccination campaign.

The prevalence of neutral threads leads to posts in neutral constituting the majority of posts in all time windows, with an average of 57% over time. Posts in negative threads make up the remaining 26% and are the ones with the greatest variance in proportion over time.

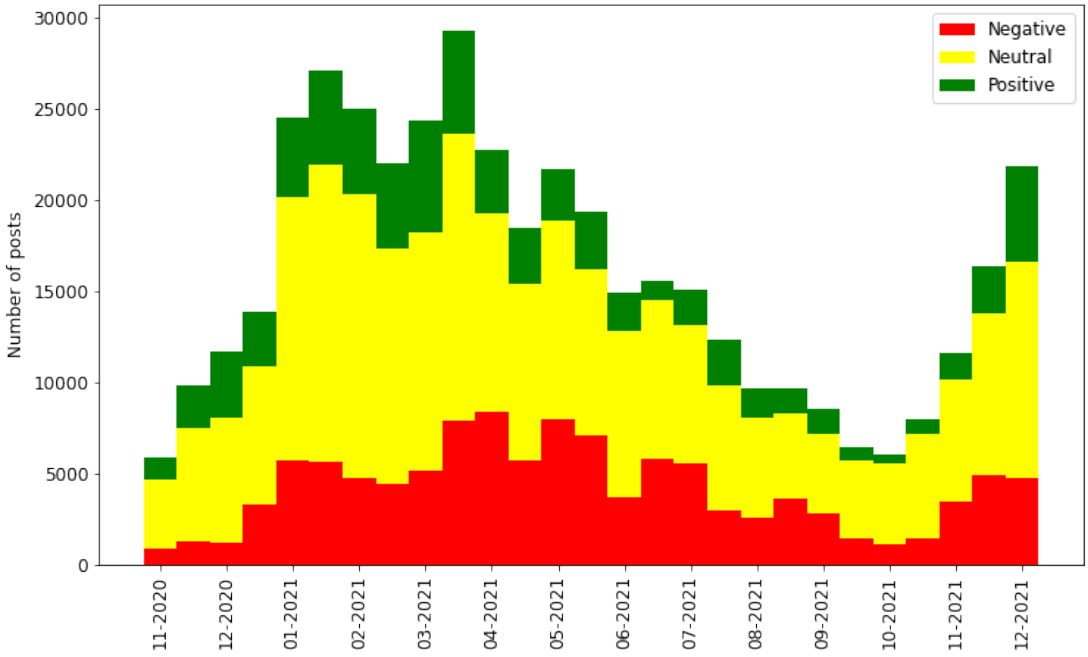


(a) Number of threads.

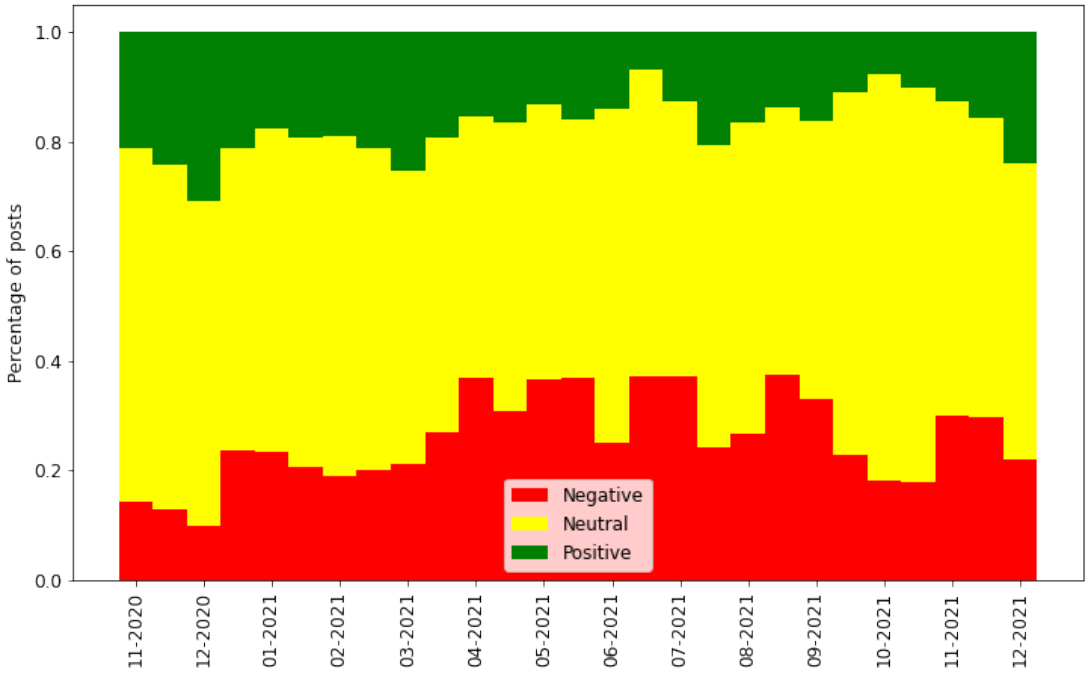


(b) Proportion of threads.

Figure 4.10: Number and proportion of threads of each sentiment label, for each data slice in the COVID-19 dataset.

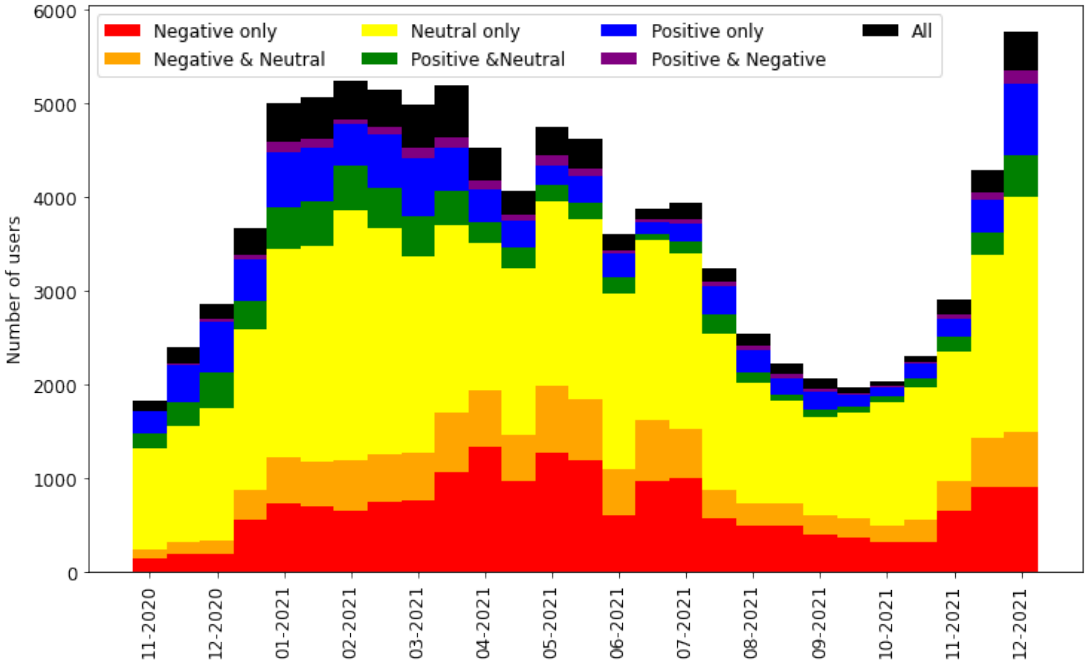


(a) Number of posts.

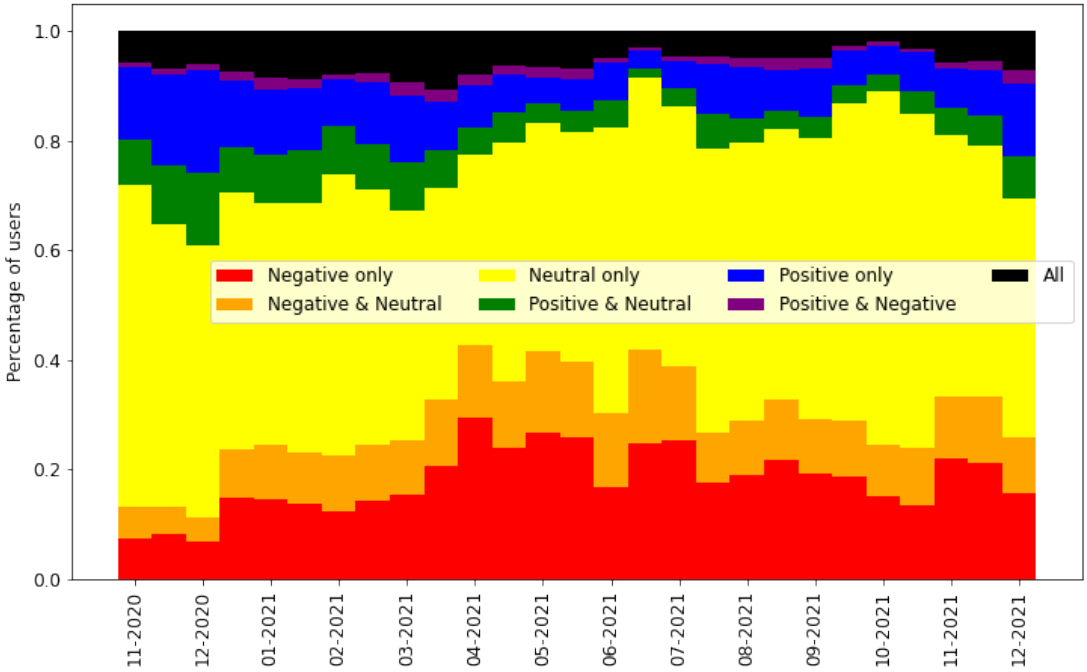


(b) Proportion of posts.

Figure 4.11: Number and proportion of posts in threads of each sentiment label, for each data slice in the COVID-19 dataset.



(a) Number of users.



(b) Proportion of users.

Figure 4.12: Number and proportion of users in threads of each sentiment label, for each data slice in the COVID-19 dataset.

In the continuation of the previous results, neutral threads are also the ones where the greatest proportion of users are found, with 48% of users posting only in neutral threads, 11% in negative and neutral, 6% in positive and neutral and also 6% in threads of all sentiments.

Figure 4.12 exposes another difference of behaviour between users discussing vaccination in the COVID-19 subforum and in the defunct vaccination subforum. Whereas in the latter we found that 89% of users restricted themselves to either positive, neutral or negative threads, in the COVID-19 dataset this number decreases to 75%. This is likely because the proportion of users that post in multiple threads is higher in the COVID-19 dataset (35% against 15% in the vaccination dataset) and they make up a larger proportion of the total number of posts (75% against 47%). In spite of this difference, when considering the users that post in multiple threads, the same percentage (69%) of these post in threads of different sentiment in both datasets.

As with the vaccination dataset, we want to correlate differences in sentiment with differences in the network structure. We previously saw that the majority of networks in the COVID-19 dataset are very similar, which fits with the results we have just exposed, as the prevailing sentiment throughout the 13 months of data is of neutrality. Nonetheless, we observe both differences in network structure and in the proportion of negative, positive and neutral threads and of posts and users therein. Using the same methodology as before, we show how the proportion of negative to positive, positive to negative and neutral to positive plus negative changes over time in regards to the number of threads (in Figure D.8), posts (Figure D.9) and users (Figure D.10).

As before, we join the network differences in 4 groups:

- The first two networks are significantly different from the remaining networks in the dataset and this difference is connected to an overrepresentation of positive sentiment in all three metrics (Figures D.8a, D.9a and D.10a) and, conversely, an underrepresentation of negative sentiment (Figures D.8b, D.9b and D.10b).

- The differences of networks starting in 15 March and 1 April 2021 are correlated with a lower proportion of neutral threads (Figure D.8c), posts in neutral threads (Figure D.9c) and users in neutral threads (Figure D.10c). The time window starting on 15 March is also the time when there is the highest proportion of users posting in multiple threads with different sentiments (Figure D.10d).

- The difference between the networks of 1 May and 1 June and between 15 May and 1 June 2021 are both above the $80th$ percentile of network distances, but

we find little evidence of sentiment change in these time periods, despite the slight underrepresentation of posts and users in neutral threads (Figures D.9c and D.10c) during the time window starting in 15 May. We also observe network differences between the networks of 15 May and 15 June and between 1 June and 1 July 2021, which can be related to the high proportion of negative sentiment in the data slices of 15 June and 1 July.

- The final networks of the COVID-19 dataset, after 1 September 2021, show a low degree of similarity, either in consecutive networks (networks starting in 15 September, 15 October and 15 November 2021) or in networks two jumps apart (networks starting 1 September, 1 October and 1 November 2021). These differences are likely to be related to a high proportion of neutral sentiment during this time, in particular the two time windows of 1 and 15 October 2021, or a low proportion of users posting in multiple threads with different sentiments.

Although we find some overlap between differences in network structure and changes in the sentiment of the forum, these do not correspond to a one to one match. For instance, the highest proportion of posts and users in positive threads occurs in 1 December 2020, but there the network structure at that point is similar to the structure of the networks that succeed it. Conversely, some network differences do not seem to be connected with changes in sentiment, such as the difference between 15 May and 1 June 2021 that we mentioned previously or the difference between 15 November and 1 December 2021.

Due to these misalignments between differences in network structure and sentiment variation, we decided to label all posts from a subset of threads from the COVID-19 forum. The threads were picked randomly, with equal proportion of positive, negative and neutral threads. Some time windows were underrepresented in the initial sample, so we re-sampled more threads from each time window that was underrepresented until there was at least 1% of threads and posts from each time window in the sample. This led to a total of 129 threads (about 3.0% of the number of threads) which contained 6898 posts (about 3.1% of the number of posts).

The posts in this dataset were labeled using the same rules as the ones we used to label the original post of each thread. Upon labeling these posts, we calculate the proportion of posts of each sentiment in threads of each sentiment. Table 4.2 shows this average proportion for each combination in the threads that we labeled. The average shown is the *macro* average, i.e., the average proportion of each thread; results do not change substantially when using the *micro* average instead.

		Threads		
		Positive	Neutral	Negative
Posts	Positive	51.7%	19.0%	22.5%
	Neutral	42.0%	76.9%	47.3%
	Negative	6.3%	4.1%	30.2%

Table 4.2: Proportion of posts of positive, neutral or negative sentiment in threads of positive, neutral or negative sentiment. The table is read as: posts of sentiment A make up X% of the posts in threads of sentiment B. For example, posts of positive sentiment make up 22.5% of the posts in threads of negative sentiment or posts of negative sentiment make up 6.3% of the posts in threads of positive sentiment.

The results from Table 4.2 show that negative threads have the widest range of sentiments, as users with opinions favourable towards vaccination often challenge negative views. On the other hand, despite their smaller frequency, positive threads are havens for users who wish to shield themselves from negativity in the forum, as negative threads are more frequent than positive ones. In positive threads, we find that the majority of posts are also positive towards vaccination, likely due to “positivity" threads that celebrate events like new vaccines being approved or milestones in the number of doses administered. Of all sentiments, neutral threads have the smallest proportion of negative posts, perhaps because often they are either question and answer threads that do not evolve into lengthy discussions and thus are not attractive to users who want to spread an anti-vaccination agenda; or they turn into discussions about issues other than vaccination, for example whether restrictions should be lifted.

Assuming that the proportions from Table 4.2 are representative for the whole population of threads in our dataset, we can infer the percentage of posts of each sentiment over time, shown in Figure 4.13.

Even though there is an overall higher amount of posts in negative threads than positive threads, a higher proportion of posts are expressing a positive sentiment than negative. We find that 26% of posts convey a positive sentiment, which is substantially greater than the 17% of posts written as replies to positive threads. Conversely, the proportion of posts in negative threads was an overestimation of the total percentage of negative posts in the dataset, which we calculate to be an average of 11% across all time slices. The proportion of posts in neutral threads was also an underestimation for the overall percentage of neutral posts, changing from 57% to 63%.

Employing the same methodology as before of Z-scoring the proportion of positive

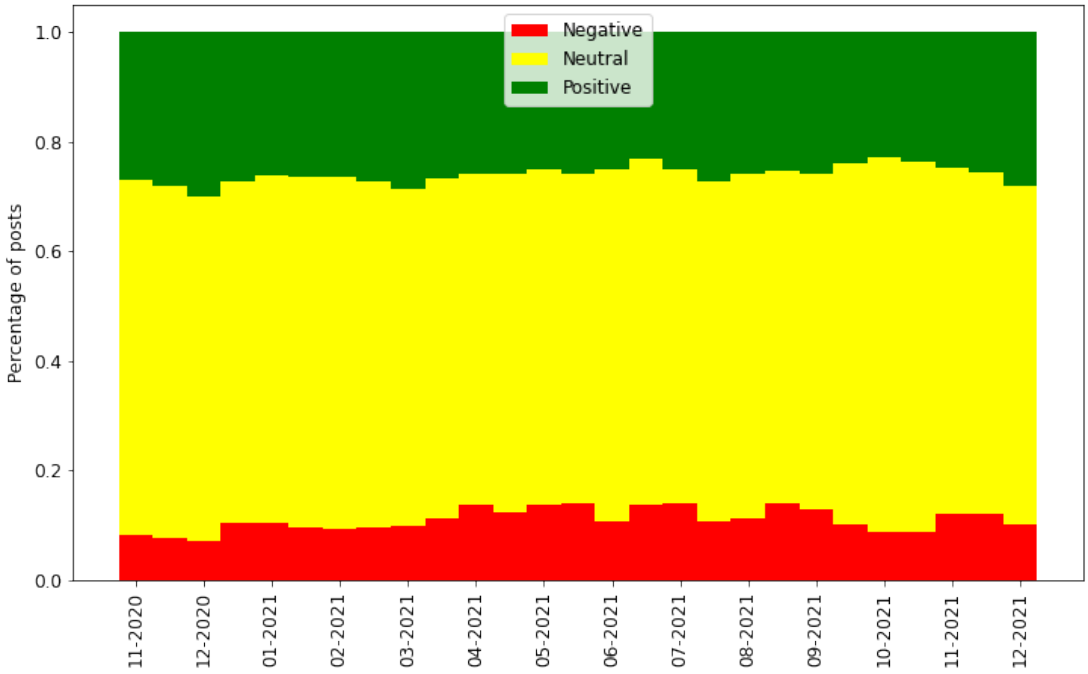


Figure 4.13: Percentage of posts of each sentiment label, inferred from the proportion of posts of each sentiment in threads of each sentiment in the labeled sample of the COVID-19 dataset, results shown for each data slice.

to negative, negative to positive and neutral to positive plus negative posts, we are able to identify a possible reason for the network structure difference between the network of 1 June 2021 to its neighbours. As can be seen in Figure D.11, in the data slices of 1 May, 15 May, 15 June and 1 July 2021 there is an underrepresentation of positive posts and an overrepresentation of negative ones, which is not the case for the time window starting in 1 June 2021. The same effect happens for the number of users in negative threads (Figure D.10b), which are overrepresented in the four closest time windows to 1 June 2021, but not in the time window of 1 June 2021.

Finally, we explore the level of disagreement in each thread of the sampled subset of the COVID-19 dataset over time, through the discordance index metric we propose. Figure 4.14 shows the average discordance per thread over time, with the discordance for each thread calculated as the average discordance using windows of 2 to 5 posts.

Results using the discordance index indicate that the network difference we observe in 1 June 2021 may be connected with a period of a high degree of disagreement within the forum. Another period of time that stands out is the particularly low level of disagreement in the time window starting on 1 October 2021, which we also previous highlight as it coincides with a significant difference in network structure. On the other

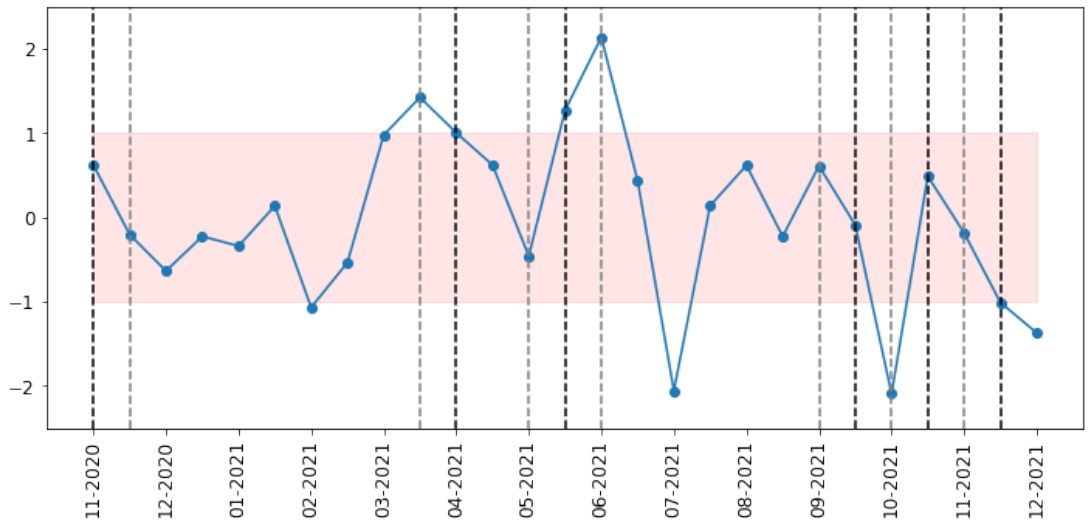


Figure 4.14: Average discordance index per thread in the sampled subset of the COVID-19 dataset over time, calculated as the average discordance using windows with 2 to 5 posts.

hand, the discordance index suggests that during the time window of 1 July 2021 there is a very low level of disagreement that is not picked up by the network comparison.

Assuming our sample is representative of the whole dataset, we can calculate the average discordance value in threads of each sentiment type and use it to infer the average discordance index for the whole dataset, which we show in Figure 4.15.

This methodology still emphasizes 1 June 2021 as a time of particularly high level of disagreement in the forum. The results also uncover a possible explanation for the network structure changes after September 2021, as the time windows starting in 1 September, 15 September and 1 October 2021 have the lowest levels of disagreement of all time frames and they are followed by a rapid increase in the discordance index in 15 October 2021.

4.5 Discussion

4.5.1 Summary of results

We have shown how user interactions with the Mumsnet forum can be represented through complex networks. By splitting these interactions according to their date, we are able to obtain snapshots of these interactions over a fixed length period of time. We employed the technique of network comparison to observe differences between each

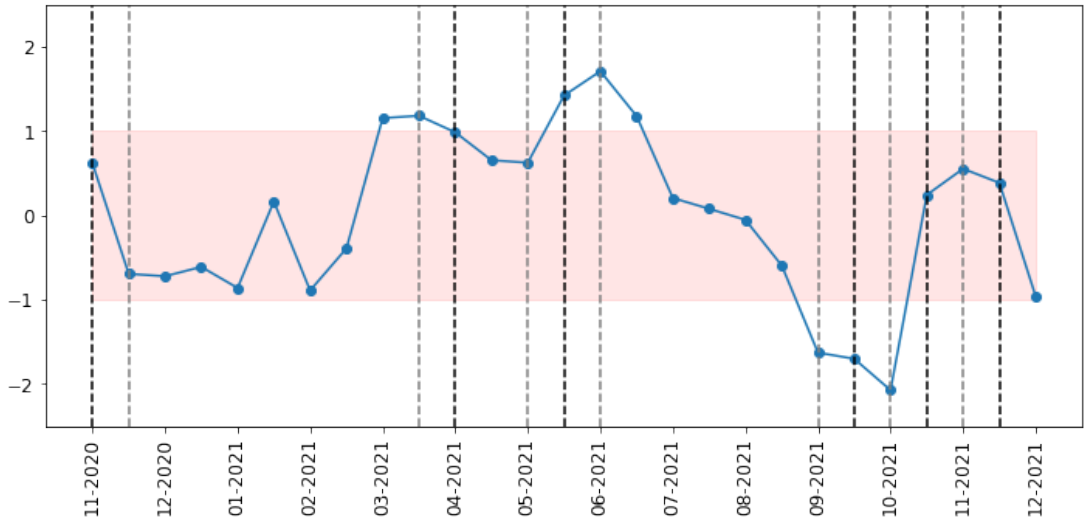


Figure 4.15: Average discordance index per thread in the COVID-19 dataset over time, calculated as the average discordance using windows with 2 to 5 posts. Values are inferred from the sample of sentiment labeled posts.

snapshot of interactions, by measuring how the structure of the network obtained from these interactions changes over time.

We found that differences in network structure often occur simultaneously to real world events that affect discussions in Mumsnet. For instance, the three major differences in the network structure of the COVID-19 forum coincide with the approval of the Pfizer and Astra Zeneca vaccines, the emergence of the Delta variant and the start of the vaccine booster program. Some of the differences in network structure we identify can be traced to changes in the distribution of specific orbits, which form a basis for the interpretation of the reasons for the differences in structure. An example of a change in the distributions of a subset of orbits that leads to significant differences in the structure of the whole network are the distributions of cliques in the vaccination subforum; these orbits are the biggest contributors for the separation of the networks in this dataset into two broad clusters, before and after May 2015.

By performing sentiment labeling and analysis on these datasets, we were also able to uncover a link between differences in the network structure and changes to the sentiment toward vaccination in the forum. Out of 10 groups of network changes, only in one we do not find any connection to an over or underrepresentation of positive, negative or neutral sentiment. Conversely, there are only two changes in sentiment that are not reflected in differences of network structure and both occur in the vaccination dataset.

Through a sample of the COVID-19 dataset, we studied whether the distribution of sentiment within the posts of a thread was a better indicator for differences in network structure. The distribution of sentiment in a thread can be seen as a proxy for the amount of disagreement within that discussion; although posts of the same sentiment does not necessarily indicate agreement, posts of different sentiment are a reflection of opposing views. By using this metric of disagreement, results show that the time with most disagreement in our sample coincides with the second largest difference in network structure.

Considering Mumsnet as a whole, our analysis unsurprisingly shows clear distinctions between how users discuss vaccination in general and vaccination within the context of COVID-19. This distinction starts in the way that discussions progress in each thread, with the COVID-19 forum characterized by short-lived threads with a high volume of replies that contrasts with the more drawn out threads, some that even last multiple years, in the vaccination forum. There are also big differences between the two forums in terms of the sentiment towards vaccination. Because the data for the vaccination forum spans a much longer period of time, it is natural that it contains greater shifts in sentiment over time. However, there is a much higher prevalence of neutral sentiment in the COVID-19 forum, as neutrality is the dominant sentiment in all time windows; the same does not happen in the vaccination forum, where there are time periods dominated by negative sentiment. Another key difference between the two forums is the amount of users posting in threads of different sentiments, which can be seen as a proxy for the amount of information flowing between threads of different types. The proportion of users posting in threads of all sentiments is significantly higher in the COVID-19 forum than in the vaccination forum, which can be interpreted as an indication that there is a wider variety of opinions in each thread of the COVID-19 forum.

Underlying all these distinctions between the two forums, we find significant differences between the network structure of their user interactions. Although the contrast in network structure is to be expected for the reasons above, we observe unexpected similarities in the distributions of certain orbits. These similarities highlight common patterns of user interactions in Mumsnet as a whole and further accentuate the difference between the networks before and after May 2015, in the vaccination forum.

4.5.2 The “so-what" factor

Sentiment labeling is a very laborious process that consumes a great deal of human resources and does not scale to large datasets, as an increase in the quantity of data requires a corresponding increase in the effort of labeling. Automatic methods have been developed but the vast majority focuses on Twitter data and are unsuitable for more nuanced arguments that do not fit the character limit of a tweet. Furthermore, automatic methods require representative samples for training and pre-labeled data for training is often inappropriate due to changes in the context in which the discussions took place.

On the other hand, the network comparison methodology we use to analyze user behaviour in vaccination discussions is easy to use as an out-of-the-box tool and scales to large social networks. Through network comparison, we were able to identify periods of time with differences in user behaviour that correlated with changes to the sentiment towards vaccination. Our methodology naturally does not replace sentiment analysis, but instead can be used as a tool to guide when it is most useful to perform sentiment analysis, possibly saving resources in the sentiment labeling.

Another strength of using our approach of network comparison is that we are able to inspect which graphlets and orbits are responsible for the differences in network structure. This knowledge can be used to compliment sentiment analysis to generate hypothesis about differences in user behaviour.

4.5.3 Limitations

Although we are able to judge which graphlets are responsible for the changes in network structure, we were unable to find whether there is a correspondence between changes in certain graphlets and the prevalence of a specific sentiment. We tried alternative techniques, such as motif analysis through the comparison against a null model or even simply inspecting the frequency of graphlets, but changes in network structure are often quite subtle; it is not the prevalence of each subgraph that leads to the differences in network structure, but rather how they are distributed over the network. This is what makes complex networks complex, characterizing their structure is often hard so it is important to develop heuristics that can help identifying when the structure is different compared to other networks that are hypothesized to be similar. Nevertheless, the relationship between network structure and sentiment is still an open problem and albeit there are other techniques we could use to uncover this relationship, we leave

that for future work.

One of the biggest limitations to the analysis of the COVID-19 forum is the wide criteria we used to include threads in our dataset, as all threads that included any keyword related to vaccination was considered in the analysis. This lead to the inclusion of threads where the discussion was not focused on the vaccines themselves but instead related to the pandemic in general. This wide criteria was particularly evident when labeling the sampled posts, where we identified discussions about lockdowns and opening up the country after the vaccination campaign, about Brexit in the context of Astra Zeneca vaccine distribution and other pandemic specific topics like modeling, masking and testing. These discussions were labeled as neutral towards vaccination but there was a large body of arguments in regard to each of these issues, with polarizing opinions between users.

It was unclear how to deal with this issue. Even if on a thread level it was possible to detect threads whose discussion was not aimed at the vaccines in particular when we labeled each original post, we found that even in threads that started off as discussions about the vaccine, the natural flow of discussion trended towards other issues. This was especially common in threads with a large number of posts, where off-topic concerns kept being brought up, usually as a reflection of the pandemic situation at the time.

These side discussions present a confounding factor to connect the structure of the networks with the sentiment of the forum. The structure of the network carries information about how the discussions are being carried out, regardless of their content, whereas all the side discussions were grouped under the umbrella sentiment of neutrality. This affects all facets of our analysis but in particular the discordance index, as we observed long sequences of neutral posts that hid a high level of disagreement.

Another issue that affects both COVID-19 and vaccination forums is moderation by Mumsnet staff in an effort to remove blatant misinformation. This is reflected in threads being deleted that we were unable to scrape or posts whose content was deleted, as well as replies that quoted those deleted posts. Whereas the former only affects the overall perception of how much negative sentiment is in the forum, as naturally only threads negative towards vaccination were deleted, the deleted posts are still encoded in the network structure and were marked as neutral sentiment. The reason to mark deleted posts as neutral instead of negative is that they do not necessarily express anti-vaccination views borne out of misinformation, but instead often they are unfruitful discussions between users that involve personal attacks and thus violate the rules of the forum.

4.6 Conclusion

This chapter presents a novel approach on the analysis of vaccination discussions in online social networks. In particular, we study how these discussions are carried out in Mumsnet, a discussion board targeted towards parents. This medium is often overlooked by researchers in favour of more popular social networks like Facebook or Twitter, but studies have found that discussion boards such as Mumsnet have a greater effect on vaccination decisions than Twitter and are a reflection of public opinion.

By recording user's posting patterns in threads, we are able to construct networks of user relationships, which represent how users behave and interact among themselves in the forum, over fixed lengths of time. One way to measure how these relationships change over time is by employing network comparison through distributions of graphlets. The methods we use to compare these networks require little parametrization, making them usable as out-of-the-box tools, scale for social networks with millions of users and are interpretable by showing which functional structures are responsible for the differences in structure.

We validate the output of the network comparison measure against indicators of sentiment towards vaccination in the forum, after hand labeling the original post of each thread related to vaccination. Results show that there is a correlation between differences in the network structure of user interactions and changes to the sentiment. Thus, network comparison can be used as a tool to reduce the effort of sentiment labeling, a labour intensive and often expensive process, by hinting when sentiment is changing and so when it might be more important to allocate resources to.

We conclude this chapter by revisiting the research questions we posed in the introduction, shedding an alternative light on the contributions presented in this chapter:

RQ1: We have shown significant differences between how users interact in Mumsnet about vaccinations in general and COVID-19 vaccinations in particular. These differences are visible not only in how often users repeatedly engage with the forum and the duration of discussion threads, but also in the sentiment towards vaccination. We have shown that the network structure of these discussions varies over time, reflecting real world events that impact discussions on the forum, and that the network structure of vaccination discussions are different from discussions about the COVID-19 vaccines. Finally, we were unable to find a distinctive pattern that characterizes either discussions about vaccination in general or COVID-19 vaccines in particular, although we found that differences in network structure were often connected to changes in the distributions of cliques.

RQ2: We have found evidence in our datasets that changes in sentiment within Mumsnet occurred at the same time as differences in network structure, both in discussions about vaccination in general and COVID-19 vaccines in particular.

Chapter 5

Contact Tracing and Testing for COVID-19

5.1 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus causing COVID-19, has continued to spread in the United Kingdom (UK) throughout 2020, 2021 and 2022 facilitated by the emergence and progressive dominance of new viral lineages such as the Alpha, Delta and Omicron variants of concern [189]. At time of writing, over 22 million confirmed cases and over 175 thousand deaths related to COVID-19 have been reported in the UK [53].

Each wave of infection has posed questions about the appropriate interventions to be made to mitigate the impacts of the disease. While the first two waves of infection in the UK were mainly managed through 'lockdown' measures involving policies such as mandated working from home and limits on out-of-household mixing, by 2021 vaccines had become available, with over 30 million first doses distributed by the end of March 2021. At time of writing, over 53 million first, over 50 million second, and over 39 million third doses have been distributed in the UK [53].

As 2021 progressed, and even prior to the arrival of the Delta and Omicron variants, there was a concern that a fast growth in cases – particularly driven by a variant either evading immunity or with enhanced inherent transmissibility – might lead to large absolute numbers of severe cases, putting pressure on an already stretched healthcare system, and leading to a large number of deaths [37]. The major legal restrictions to economic, educational and social activities that had been implemented as emergency measures to disrupt transmission during the previous 18 months, which were believed

to be of increasing harm [177] and decreasing efficacy over time [32], provided a strong motivation in late 2021 to find ways of reducing SARS-CoV-2 transmission using other methods.

An alternative targeted approach to ongoing lockdowns and restrictions on social gatherings, is seeking to identify those who are infectious and restricting isolation to only those individuals. This is the goal of Test-Trace-Isolate (TTI) strategies, but using TTI alone to control growth was found to be challenging early in the pandemic [30, 40, 46]. Identifying individuals who are infectious before they have the opportunity to make contacts is difficult because many cases are only mildly symptomatic or asymptomatic, and presymptomatic transmission is common, making symptom-based case ascertainment unreliable. In the UK, from the start of the pandemic, the government invested heavily in Polymerase Chain Reaction (PCR) testing infrastructure, which was primarily used in the community to detect infections among individuals experiencing one of a set of symptoms. PCR testing involves taking samples with swabs (of the naso-pharynx, as advised by test providers in the UK), and transporting the samples to one of a network of laboratories where the sample is amplified to detect any fragments of SARS-CoV-2 RNA. Later in the pandemic life cycle, an alternative technology provided by Lateral Flow Tests (LFTs) – alternatively called Rapid Antigen Tests (RATs) or Lateral Flow Device (LFD) tests – was widely deployed in the UK. These tests are inexpensive cartridges that deliver results within 15-30 minutes, rather than in 1-2 days for PCR tests, and do not require laboratories with their associated staff or logistic networks.

While it is possible to deploy PCR at scale – as has been done most prominently in China [95] – the speed and low cost of LFTs makes these more attractive for large population testing programmes. Further, Larremore et al. [89] argue that the key criterion for a test being useful for these purposes is sensitivity during the infectious period rather than sensitivity to any level of detectable virus. According to this criterion, LFTs are not expected to be substantially inferior to PCR [81]. In high prevalence settings ($\geq 5\%$), the World Health Organisation (WHO) recommends prioritisation of LFTs as they can be quickly produced in large quantities and the results can be obtained rapidly on site and without the need for a laboratory [188]. A positive LFT result will then allow immediate isolation of suspected cases and timely contact tracing.

By providing a rapid result and reducing the need for resource-intensive laboratory and transport infrastructure subject to capacity limitations, LFTs offered the opportunity to expand case finding strategies to asymptomatic as well as symptomatic

individuals on a wider community level. Mass use of single LFTs targeted to asymptomatic individuals or regardless of symptoms had a significant impact when deployed nationally in Slovakia [126] and locally in Liverpool [97], although these schemes also required significant human resources in their delivery. Through the first half of 2021, the UK applied setting-specific asymptomatic testing programmes, such as in care homes, healthcare settings, schools and workplaces and some evaluation of the potential efficacy and effectiveness of these programmes has occurred [18, 90, 195].

At the start of Autumn 2021, in order to avoid both the direct harms of a large Winter epidemic wave and implications for healthcare capacity, as well as the indirect harms of restrictive lockdown measures and social contact restrictions, there was a need to assess the potential effectiveness of a regular population-wide asymptomatic screening programme using LFD tests. Our work aims to explore variations on such a policy, against the background of the existing symptomatic testing and tracing programme, using an adaptation of a published agent-based model (ABM) called *Covasim* [82]. Covasim was previously used in the UK to evaluate the impact of different TTI strategies when schools reopened after the first national lockdown [122], to explore the impact of wearing face coverings in schools [123], to simulate different lifting lockdown lifting strategies [124, 125], and as part of the UK Health Security Agency's epidemic nowcasting [52].

Covasim is a network based model, incorporating multiple networks that represent settings where infection-transmitting contact can occur. The lockdowns introduced to fight the spread of COVID-19 have had an impact on social contact patterns of the whole population. Policies such as social distance, school closures and advice to work from home led to cautious behaviour from the population in general and social contact surveys show that, by March 2022, the average number of daily contacts still had not reached the same values as before the start of the pandemic [77]. Using age-mixing matrices from social contact surveys before [111] and during [78] the pandemic, we extend Covasim to generate contact networks that reflect changes observed in these surveys. By applying network comparison to the networks generated from the age-mixing matrices, we can gain understanding about the different infection-spreading patterns that emerge from the age-mixing patterns of different time periods.

Against this backdrop of different social contact patterns, this chapter extends existing Covasim studies to consider different asymptomatic testing regimes and choice of testing technology for symptomatic testing, across a range of adherence levels and epidemic growth rates. Specifically, we answer the following five questions:

1. What is the effectiveness of regular LFD testing among asymptomatic individuals in reducing the total number of infections and the size of the peak, when combined with a moderate (40%) background PCR testing of symptomatic individuals combined with PCR contacts testing under scenarios:
    - According to asymptomatic testing frequency
    - For different levels of asymptomatic testing and isolation uptake
2. What percentage of non-household contact reduction would be required to obtain equivalent effectiveness in reducing peak size/number of infections compared to a regular asymptomatic testing policy?
3. If there is no asymptomatic testing, how does the effect on number of cases/peak size of symptomatic isolation and testing, plus PCR testing of identified case contacts, vary by uptake amongst those symptomatic and their traced contacts?
4. What are the implications for testing with a single LFD test compared to a single PCR test when symptomatic?
5. For the range of testing policies, what are the associated 'costs' in terms of numbers of days people are asked to isolate at home and numbers of tests required?

5.2 Methods

5.2.1 Model Overview

For the purposes of this analysis we used a modified version of Covasim based on release 3.0.2. The modified Covasim and the code for running all simulations contained in this paper are available at `https://github.com/TTI-modelling/covasim` and `https://github.com/migueleps/RegularAsymptomaticTesting`.

Treating each individual as an agent, *Covasim* incorporates individual transmission variability, age-specific disease progression probabilities, immunity dynamics, and the effect of various vaccination and non-pharmaceutical interventions. Each timestep in the model represents one day and is split into three main phases: interventions, immunity and transmission. Interventions include both non-pharmaceutical interventions, like contact tracing, and pharmaceutical ones, like vaccination. Immunity in the model is implemented through Neutralising Antibodies (NAbs) trajectories, which are mapped to reductions in transmissibility, susceptibility and disease severity, a process

described in [23]. Transmission is calculated on a per connection basis, controlled by a parameter β representing the transmission probability. Thus, each contact between a susceptible and an infectious agent has an independent probability β of leading to a transmission event. However, this parameter is governed by multiplicative factors that affect the transmission probability, for example, the contact layer where the contact occurred or other factors, such as age-based susceptibility. In our analysis, we used the majority of default parameters regarding transmission and disease progression in *Covasim*, while the parameters specific to our simulations and studied interventions are documented in Appendix F. As suggested by Kerr et al. [82], we calibrate the parameter β to produce epidemics with desired initial growth rates.

In addition to the transmission probability varying according to the contact layer and other susceptibilities, we also vary the transmission probability per individual and over time. Infectiousness of an individual is expected to be directly dependent on their viral load. However, it is difficult to obtain data on the distribution of viral load trajectories in general populations and the relationship between viral load and infectiousness is not known. As such, we find it necessary to take the simplifying assumption that the infectiousness of agents depends solely on their infectious age. We base our model of infectiousness upon the shape of the generation time distribution, defined as the distribution of times from infection to transmission, estimated by Ferretti et al. [39] who combined several different datasets of infector-contact pairs to produce an estimate for the generation time distribution.

5.2.2 Contact Networks

We simulated a population representative of a UK town, which consists of 100,000 individual agents linked by a 4-layered contact network (household, school, work, and community). The contact structure (i.e. the average number of contacts per person) and the transmission probability per contact are different in each layer. In the household layer, the average number of contacts per person is drawn from a Poisson distribution with mean of 2.25, informed by the mean household size in the United Kingdom [174]. All members of a household are fully connected to the remaining members of that household and disconnected in this layer from any person outside the household.

We parameterized the structure of the remaining 3 contact layers using age-dependent contact matrices obtained from the social contact surveys CoMix [78] and POLYMOD [111]. The UK CoMix survey is conducted bi-weekly in real time since March 2020, which records the dynamics of individual's contact behaviours alongside

the pandemic, while POLYMOD is the first comprehensive contact survey in European countries in 2008, which represents pre-pandemic contact patterns. POLYMOD contact matrices for the UK are available publicly online and we extracted contact matrices from Rounds 22 to 27 (late August 2020 to September 2020) of CoMix survey as they documented the highest average number of contacts since the start of COVID-19 pandemic, as of November 2021.

The CoMix surveys were designed to be consistent with POLYMOD, recording age of the participant, age of contact and the setting the contact occurred in. The ages of participants and individual contacts are split into 9 age bins: 0 to 4 years, 5 to 11 years, 12 to 17, 18 to 29, 30 to 39, 40 to 49, 50 to 59, 60 to 69 and 70+ years. The CoMix survey asks participants to make a distinction between individual contacts and "group contacts" [51], a category for participants who are unable to recall details for all individual contacts (for example, those working in public facing jobs). This category is split into less fine-grained age divisions: 0 to 17, 18 to 64 and 65+ years. These age bins are matched to the more detailed ones by sampling uniformly at random an age bin in the appropriate range, for instance, to assign a contact from the 0 to 17 age bin, we sample uniformly at random from the 0 to 4, 5 to 11 and 12 to 17 age bins. Following the analysis on the weekly CoMix reports, we truncate the maximum number of contacts by each individual to 50 per day.

Using these age-mixing contact matrices, we generate each contact network layer using a configuration model. The number of contacts of each individual to each age bin is drawn from a negative binomial distribution.

Unlike the household, school and work layers, the community layer uses Covasim's dynamic layer setting [82], wherein contacts are regenerated after each timestep, while preserving the number of edges between age bins.

5.2.3 Interventions

We modeled government policy in our simulation setup through the inclusion of four interventions: vaccination; isolation on symptoms; testing; and contact tracing.

5.2.3.1 Vaccination

We assume a vaccination coverage scenario where 90% of people aged over 40 have been vaccinated with 2 doses of the Astra Zeneca vaccine and 45% of people aged

under 40 have been vaccinated with 2 doses of the Pfizer vaccine. This leads to approximately 67% of the population being vaccinated, which at the time we started this work, was considered a high vaccination coverage scenario, but it is likely to be underestimating population coverage at the time of writing. We also assume that the whole population is vaccinated from day 0, so the decaying vaccine efficacy resulting from decaying neutralizing antibodies is the same for all age groups. Through previous simulations, we measured that this vaccination scenario reduces the number of total infections by 70% in an epidemic with growth rate 0.15, equivalent to a doubling time of 5 days. This vaccination strategy is not meant to reflect the progression of the vaccination effort observed in the UK. Instead, it was chosen to consider the pre-existing population immunity at the start of this work, following the questions posed to us by SPI-M.

5.2.3.2 Isolation on symptoms

Until April 2021, UK government policy advised that, upon developing at least one of three COVID symptoms (fever, new continuous cough or loss of smell or taste), people should self-isolate and request a PCR test, remaining in isolation until either receiving a negative result or for an isolation period following a positive test (which varied, but was 10 days for the majority of pandemic during 2020-2021). However, there was variation among the population in the ability to identify COVID symptoms correctly or to isolate prior to receiving test results. To take this into account, we implement isolation on symptoms with an adherence parameter, the probability that someone will preemptively isolate when developing symptoms.

As a respiratory disease, COVID shares symptoms with Influenza-like Illnesses (ILI), which can lead to people mistaking an ILI infection with COVID and preemptively self-isolate while waiting for a test result, despite not being infected with COVID. We model this scenario by picking randomly 1% of the population to be infected with ILI symptoms every day and self-isolate with the same adherence probability.

Both groups of people isolating on symptoms are released from isolation if they test negative at least one day after developing symptoms.

5.2.3.3 Testing

Testing interventions allow the identification of infected individuals within a simulation. A diagnosed individual is put in isolation, which is implemented as a reduction to

β by 90% in non-household contact layers. This isolation is always mandatory, regardless of the isolation on symptoms policy. Testing also unlocks contact tracing, which is implemented as a separate intervention but requires testing interventions to work.

We model testing differently depending on symptomatic status. In the symptomatic case, people with symptoms for COVID or ILI have a fixed probability of testing every day. Those who test positive are placed in at-home isolation for 10 days, regardless of their isolation status prior to the test.

In the asymptomatic case, we split the population in three groups: regular testers, intermittent testers and anti testers. Regular testers are the ones that follow government advice rigorously, testing at periodic intervals. The rest of the population is split in 10% for the anti testers, who never test if they are asymptomatic, and the remaining 90% for intermittent testers, who test at a quarter of the rate of the regular testers. This means that if 60% of the population is testing regularly twice per week, then 6% of the population never test if they are asymptomatic and 34% of the population will test on average once every two weeks (7% probability of testing per day). These parameters were chosen as a best guess following conversations with specialists, as specific parameters for modelling are hard to estimate during an evolving pandemic [103].

As testing time is known as an important feature impacting the effectiveness of TTI interventions, we added an extension to *Covasim* which incorporates time depended individual PCR and LFD test sensitivity trajectories estimated from empirical regular testing of healthcare workers data [60]. We also include a testing delay for PCR tests, given by a Poisson distribution with mean 1.2, a swab error rate of 10% (tests that return negative regardless of infection status) and a LFD test specificity of 99.7%.

5.2.3.4 Contact tracing

A positive test through a testing intervention triggers contact tracing. A proportion of contacts of the diagnosed individual are notified as contacts. The proportion of contacts and the time it takes to find them depends on the contact layer that the contact occurred in: in the household layer, we assume that all contacts are found within the same day of diagnosis; in the school and work layers, we assume that 45% of contacts are found the day after the diagnosis; finally, in the community layer, we assume that only 15% of contacts are found, taking 2 days from the case's test result to do so.

We modified the default contact tracing interventions implemented in Covasim to better reflect UK policy during Autumn 2021. Instead of asking contacts of a diagnosed positive case to quarantine, they are asked to take a PCR test instead, which they do

with probability controlled by an uptake parameter.

5.2.4 Outcomes

We focus the analysis of our model on the following outcomes:

- **Number of infections -** Total number of people who transition from a susceptible state to exposed, including reinfections.
- **Height of new infections peak -** Maximum number of new infections in a single day.
- **Number of tests -** Average number of tests, either PCR or LFD, that each individual in the simulation performs during the 180 days of simulation.
- **Person days of isolation -** Average number of days that each individual is in isolation during the 180 days of simulation.

5.2.5 Summary of changes to Covasim

To highlight the contributions towards Covasim that underpins our experimental setup, we provide here a summary of the changes we implemented in Covasim:

- Contact networks: we implemented a generation mechanism that takes as input an age mixing matrix to generate networks with age contact patterns similar to those observed in POLYMOD and CoMix.
- Interventions: we modified how testing works to include false positive tests and bad swabs, as well as the testing scheme we discussed in Section 5.2.3.3; we reworked the implementation of test sensitivity, so that it is different for PCR and LFD tests and depends on time since infection, instead of a constant probability of a false negative; we introduced a new intervention to model isolation on symptoms and ILI infections, including an alteration in how isolation is processed; and we changed contact tracing to no longer induce a period of quarantine but instead to trigger a testing event.
- Analyzers: we added new analyzers to track person days of isolation for each agent in the simulation.

5.3 Results

We split our experiments into three sets of simulations, designed to answer the five research questions we proposed. Each set of simulations was run with two contact patterns, one obtained from POLYMOD and the other encompassing the CoMix survey rounds 22 to 27, referring to late August 2020 and September 2020, the time period when the number of average contacts was highest from the beginning of the pandemic until November 2021, when the simulations were executed.

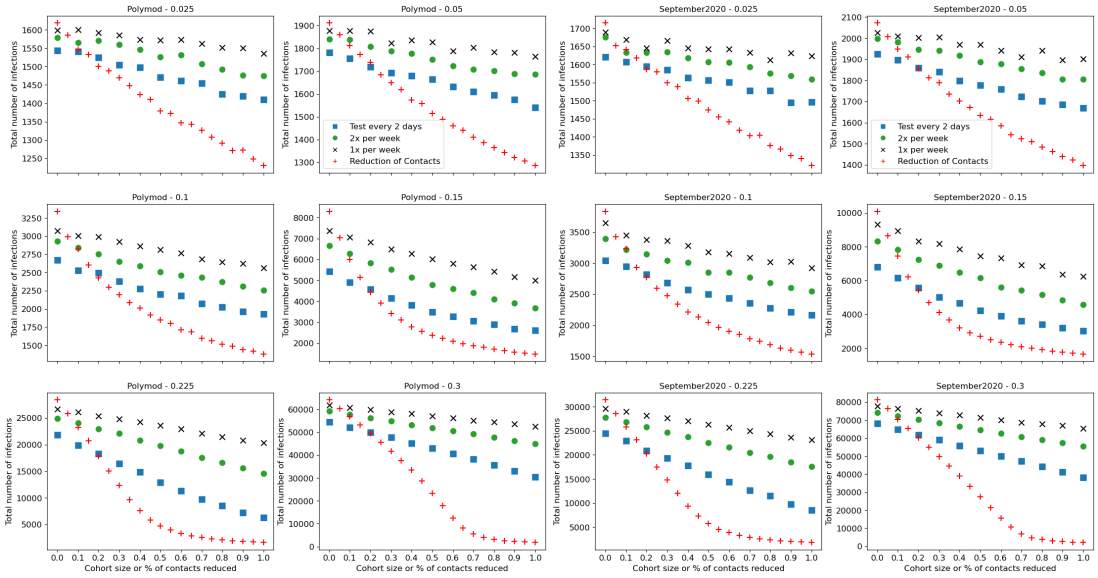
We also calibrate the β parameter per infectious contact to multiple growth rates, for each of the contact patterns. The growth rates we use are 0.025, 0.05, 0.1, 0.15, 0.225 and 0.3, which correspond to approximate doubling times of 1 month, 2 weeks, 1 week, 5 days, 3.5 days and 2.5 days, respectively.

5.3.1 Asymptomatic testing and reducing number of contacts

To study the impact of increasing asymptomatic testing, we vary the percentage of the population in the regular testing cohort from 0% to 100%, in intervals of 10%. We also vary the frequency of testing, with the regular groups testing once per week, twice per week or every two days. The intermittent testers have a daily probability of testing of 1/28, 1/14 and 1/7, respectively, meaning that they will test on average once a month, once every two weeks or once per week. In this set of simulations, we fix the adherence to isolation on symptoms, the daily probability of testing if symptomatic and the proportion of contacts that take up a PCR test to 40%.

Government policy such as social distancing and sector closure ('lockdowns') were introduced to prevent transmission of COVID by reducing the number of contacts each person makes. To represent these interventions in Covasim, we randomly remove a percentage of edges in the work, school and community layers of the contact network (i.e. not the household network). We vary the percentage of edges we remove from the network in increments of 5%, from 0% to 100%. For the simulations in this scenario, we remove all asymptomatic testing and set adherence to isolation on symptoms, daily probability of testing if symptomatic and the proportion of contacts that take up a PCR all to the same value as before (40%), for a comparison of the impact of reducing contacts against increasing asymptomatic testing.

We show the comparison between these two approaches through the number of total infections after the 180 days of simulation, in Figure 5.1, and the number of person days of isolation, in Figure 5.2. We also show the relationship between the



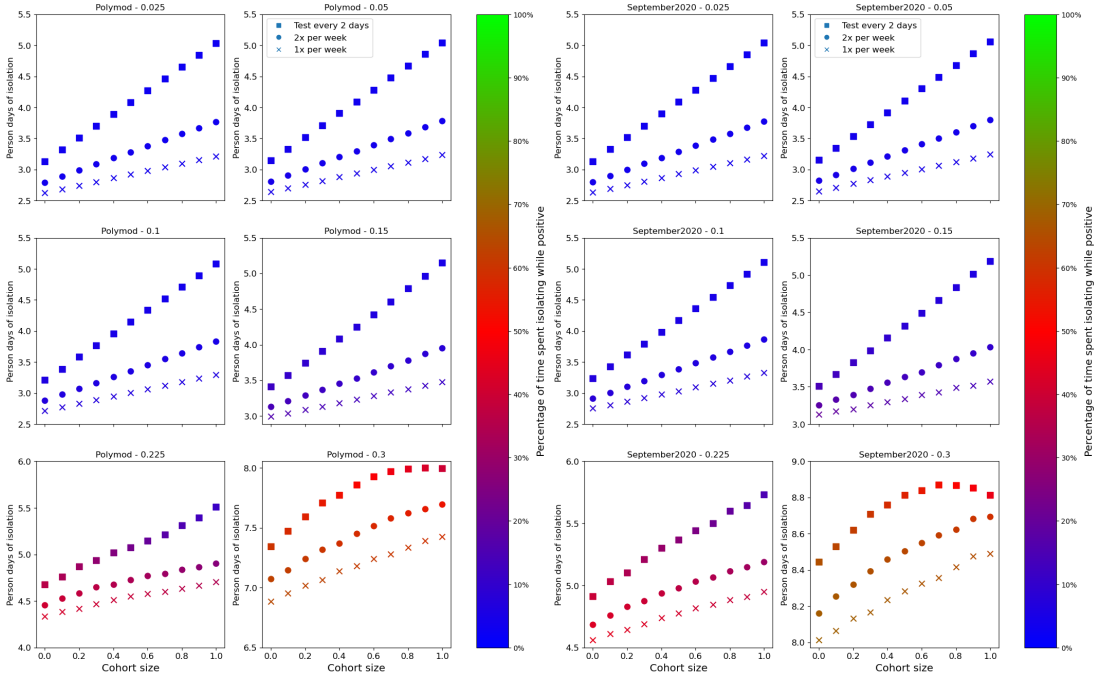
(a) POLYMOD contact patterns (b) CoMix contact patterns

Figure 5.1: Comparison of asymptomatic testing strategies against reduction of contacts for different growth rates and contact patterns. The x-axis represents the size of the regular tester cohort for the simulations with asymptomatic testing or the percentage of contacts removed from non-household layers for the simulations with contact reduction. The y-axis shows the total number of infections, averaged over 100 runs.

size of the peak in number of daily infections and the number of total infections, in Appendix G (Figures G.1 and G.2).

For a twice weekly testing policy, which was the government advice for regular asymptomatic testing during 2021, we find that at 100% adherence, the decrease in total number of infections compared to 0% adherence to this policy is within the range [6.6%, 44.6%] for POLYMOD contact patterns and [7.0%, 44.9%] for CoMix contact patterns, across the different growth rates, Figure 5.1. The growth rates at which these boundary values occur are 0.025 and 0.15 for both POLYMOD and CoMix. This reduction in the number of total infections is similar to the impact of reducing the number of contacts by 25 to 30% across all growth rates, for both sets of contact patterns. At 50% adherence, these intervals are reduced to [3.3%, 27.9%] for POLYMOD and [4.1%, 26.0%] for CoMix.

For a once weekly testing policy, if 100% adherence, the decrease in total number of infections ranges from 3.9% at 0.025 growth rate to 32.2% at 0.15 with POLYMOD contact patterns, and from 3.9% at 0.025 growth rate to 32.9% at 0.15 with CoMix contact patterns. The decline in the total number of infections is equivalent to reducing



(a) POLYMOD contact patterns (b) CoMix contact patterns

Figure 5.2: Impact on person-days of isolation of increasing adherence to the different asymptomatic testing strategies, for different growth rates and contact patterns. The x-axis represents the size of the regular tester cohort for the simulations with asymptomatic testing or the percentage of contacts removed from non-household layers for the simulations with contact reduction. The y-axis shows the mean number of person days of isolation per population members over the 180 days of simulation, averaged over 100 runs. These include days in isolation waiting for test results (amongst true positives and true negatives); isolation days amongst those testing positive who were true positive; and isolation days among those testing positive who were false positives. The colour of each marker indicates the percentage of days spent in isolation while infected.

the number of contacts by 15 to 20%, across all growth rates and for both sets of contact patterns. At a lower level of adherence, 50%, this decrease is reduced to 1.6% at 0.025 growth rate and 18.4% at 0.15 with POLYMOD contact patterns, and to 2.7% at 0.05 growth rate and 19.9% at 0.15 with CoMix contact patterns.

For the most intensive asymptomatic testing policy, testing every 2 days, given 100% adherence, the decrease in total number of infections is within the range [8.6%, 71.0%] for POLYMOD contact patterns and [7.7%, 65.2%] for CoMix contact patterns. The growth rates at which these boundary values occur are 0.025 and 0.15 for both contact patterns. This reduction is equivalent to the impact of reducing the number of contacts by 40 to 45% for all growth rates, in both sets of contact patterns. At 50% adherence, these intervals are reduced to [4.6%, 40.9%] for POLYMOD and [4.0%, 37.9%] for CoMix.

We find that the impact of testing on the maximum number of infections per day is approximately the same as the impact in number of total infections, when compared to reducing the number of contacts (Figures G.1 and G.2). Results show that testing every 2 days yields a peak height equivalent to reducing the number of contacts by 40 to 45%, testing twice per week is equivalent to reducing contacts by 20 to 25% and testing once a week is equivalent to reducing contacts by 10 to 15%. These numbers are consistent across multiple growth rates and contact patterns.

Analysis of person-days of isolation reveals similar behaviour when the growth rate is less than or equal to 0.1, for both contact patterns, Figure 5.2. In these low prevalence cases, increasing the adherence to the testing policy leads to a increase of 22.5% (0.6 days), 35% (1 day) and 60% (1.9 days) in number of person-days of isolation for the weekly, twice weekly and testing every two days policies, respectively. The time spent in isolation while positive accounts for less than 10% of the time in isolation, for 0.025, 0.05 and 0.1 growth rates.

For higher growth rates, increasing testing leads to a smaller increase in the number of person-days of isolation, with the increase becoming smaller the higher the growth rate is. For the testing every two days policy, increasing adherence to 100% leads to an increase in number of person days of isolation of 51% (1.7 days), 17.7% (0.9 days) and 8.9% (0.6 days) for 0.15, 0.225 and 0.3 growth rates, respectively, when using POLYMOD contact patterns, and 47% (1.6 days), 16.6% (0.8 days) and 4.3% (0.4 days), when using CoMix contact patterns. For the twice weekly testing policy, increasing adherence to 100% leads to an increase in number of person days of isolation of 26%

(0.8 days), 10.1% (0.4 days) and 8.8% (0.6 days) for the same growth rates, when using POLYMOD contact patterns, and 23.9% (0.8 days), 10.8% (0.5 days) and 6.5% (0.5 days), when using CoMix contact patterns.

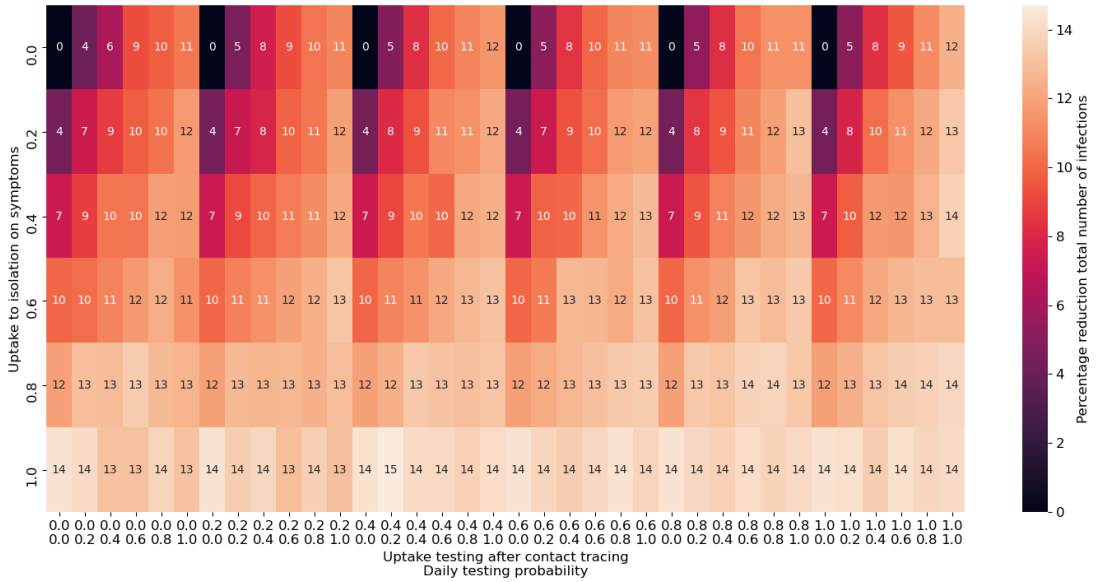
For all growth rates, this increase in isolation time is driven by an increase in people isolating from false positive LFD tests. This is supported by inspecting the proportion of time spent isolating while positive to COVID, which decreases as adherence to testing increases, represented by the colour gradient in Figure 5.2. This effect is more pronounced at high growth rates and higher rates of testing. For instance, at a growth rate of 0.3 with POLYMOD contact patterns, for the testing every two days policy, the time spent isolating while positive varies from 60% at 0% adherence to 42% at 100%, but for the twice weekly testing policy, these values change to 63% and 54%, respectively.

5.3.2 Symptomatic testing

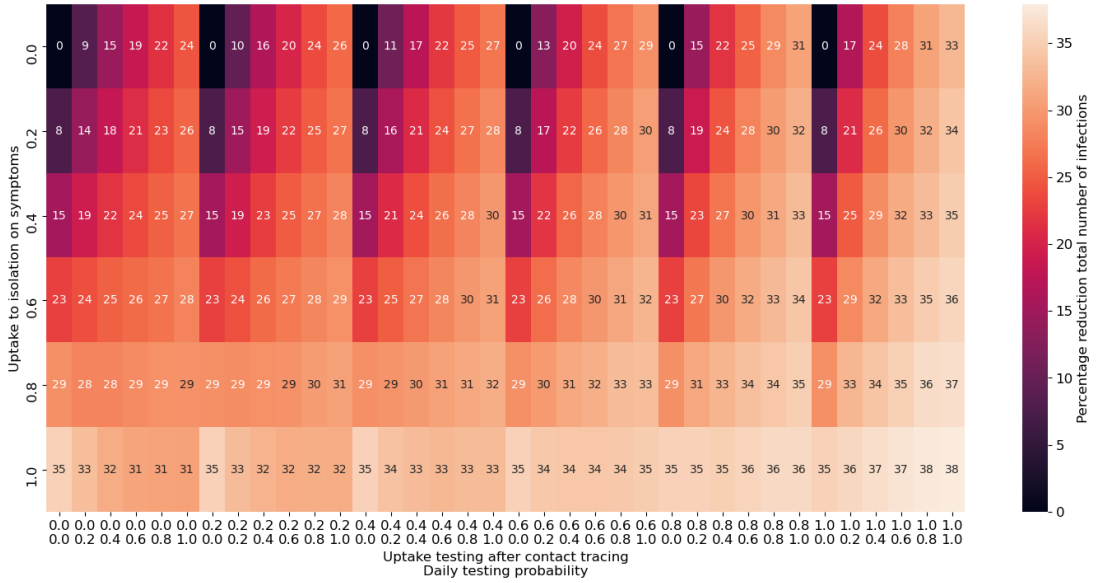
We also explore how sensitive our model is to changes in the parameters that control adherence to isolation on symptoms, daily probability of testing if symptomatic and the proportion of contacts that take up a PCR. We vary these parameters in increments of 20% from 0% to 100%. An uptake of 100% in these interventions means that all traced contacts take a PCR test when they are traced and everyone that develops COVID or ILI symptoms isolates and performs a PCR test the day after isolating, continuing to isolate for a total of 10 days if the test is positive and being released from isolation if the test is negative. These experiments are conducted solely on POLYMOD contact patterns, for the multiple growth rates as above, due to the heavy computational load. The impact of changing these uptake parameters in the total number of infections is shown in Figure 5.3.

Results show that the most effective of three interventions at reducing transmission is isolation on symptom onset, across all growth rates. This is a very intuitive result and follows our expectation, as isolation naturally prevents most transmission. Perfect adherence to isolation on symptoms with no contact tracing and no testing accounts for a reduction in number of cases between 14.3% (at a 0.025 growth rate) and 66.9% (at a growth rate of 0.15).

On the other hand, contact tracing and asking contacts to take a PCR test, isolating only if positive, is the least effective of these interventions. With a 100% testing probability when symptomatic and no isolation on symptoms, the combination of parameters that should favour contact tracing the most as contacts can only be traced when there



(a) Growth rate = 0.025



(b) Growth rate = 0.3

Figure 5.3: Impact of increasing the uptake on isolation on symptoms, daily symptomatic testing and PCR testing after being traced as a contact of a positive case. Numbers shown are the percentage decrease in the total number of infections, compared to 0% uptake on these interventions, for 0.025 and 0.3 growth rates in POLYMOD contact pattern. The top line on the x axis represents the uptake to testing after contact tracing, the bottom line represents the daily probability to testing when symptomatic. The y axis represents uptake to isolation on COVID or ILI symptoms.

is testing, our results show that contact tracing reduces the number of cases between 1.9% (at a growth rate of 0.025) and 14.5% (at a growth rate of 0.225).

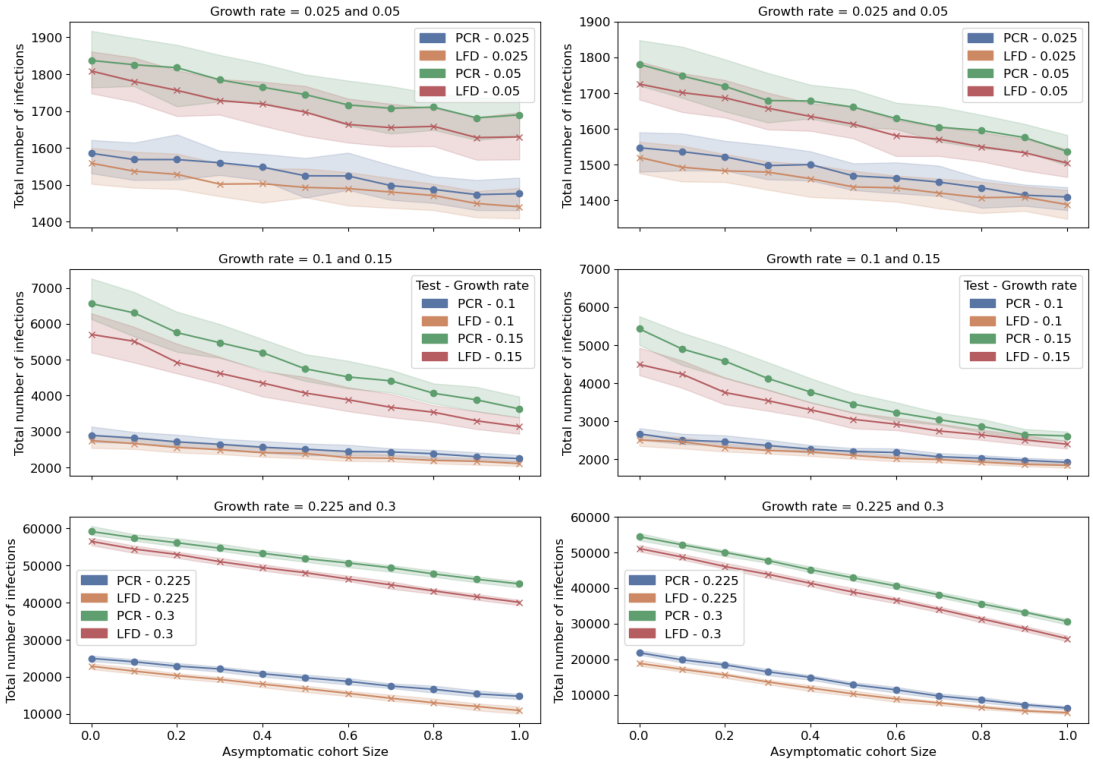
The impact of symptomatic and contact testing is harder to measure, as it interacts with the effectiveness of the other two interventions. If we assume testing but no isolation on symptoms (only after a positive test) and no contact tracing, testing can reduce the number of cases from 10.6% (at a growth rate of 0.025) to 51.6% (at a growth rate of 0.15). However, if we assume perfect adherence to isolation on symptoms, the impact of testing depends on the adherence to testing after being contact traced. For example, at a growth rate of 0.15, with no contact tracing, increasing the daily probability of testing from 0 to 100% leads to more cases, from an average of 5340 to 6235, corresponding to a reduction of 66.9% to 61.3% when compared to no interventions. But if we add contact tracing with perfect adherence, increasing the testing probability leads to less cases, from an average of 5340 to 5036, corresponding to a reduction of 66.9% to 68.7% when compared to no interventions.

The value of 40% adherence to the 3 interventions we chose for the asymptomatic experiments, correspond to a reduction in number of cases by 10.5%, 15.6%, 39.3%, 50.2%, 29.6% and 23.9%, for each growth rate, respectively. These figures also indicate that testing is most effective when the epidemic is at a medium level of growth; if the epidemic is at a very high rate of growth testing is not enough to curb transmission.

5.3.3 Comparing PCR and LFD tests

We now explore how changing all PCR tests - taken by symptomatic individuals and by contacts of identified cases - for a single LFD test impacts the total number of infections, visualized in Figures 5.4 and G.5 for the asymptomatic testing experiments and in Figures G.6 and G.7 for the symptomatic testing experiments. These results all use POLYMOD contact patterns.

While varying asymptomatic testing, we find that using a single LFD test in lieu of a PCR test among symptomatic individuals and traced contacts yields a smaller total number of infections, with particular significance at high growth rates of the epidemic. For a testing every two days policy, the difference ranges from 1.6% (0.025 growth rate) to 20% (0.225 growth rate) when the whole population is testing frequently and from 1.9% (0.025 growth rate) to 19% (0.225 growth rate) when half the population is the regular testing group. For a twice weekly testing policy, the difference ranges from 1.9% (0.025 growth rate) to 25% (0.225 growth rate), when the whole population is testing regularly, and from 1.7% to 16%, at the same growth rates, when half the



(a) Asymptomatic testing twice per week (b) Asymptomatic testing every two days

Figure 5.4: Comparison of LFD against PCR tests for symptomatic or contact traced individuals, for different values of asymptomatic testing uptake and different growth rates, with contact patterns obtained from POLYMOD. The y-axis shows the total number of infections. The x-axis represents the size of the regular asymptomatic tester cohort. Markers shown represent the median of 100 simulations and the shaded area represents the interquartile range. Figure 5.4a corresponds to the twice weekly testing policy and Figure 5.4b to the testing every two days policy.

population is the regular testing cohort.

In the experiments with no asymptomatic testing, we find that at small growth rates, there is no significant difference between using a PCR versus single LFD test for symptomatic individuals and contacts. We find that at medium growth rates (0.1 and 0.15), across different values for adherence to isolation on symptom onset and daily probability of testing, using LFD tests leads to a smaller number of infections if the adherence to testing after contact tracing is less than 40%, after which PCR leads to on average to lesser number of infections. Finally, results show that at high growth rates (0.225 and 0.3), LFD tests consistently lead to a smaller number of infections, in particular when the probability of testing if symptomatic is 100%, the average (across the range of uptakes to testing after contact tracing) difference between PCR and LFD tests is 7.6%, 6.6% and 5.0% with no isolation on symptoms, 40% adherence and 100% adherence, respectively.

Another metric of interest in this context is the number of tests required to implement this increase in adherence to testing, symptomatic or asymptomatic. In the asymptomatic case, we find no difference in the total number of tests required between using PCR or LFD tests for symptomatic individuals and contacts, which is expected because the majority of testing is done in asymptomatic individuals. Implementing the asymptomatic strategies if the whole population took up asymptomatic testing requires 30.1, 49.2 and 91.9 tests per person for the weekly, twice weekly and testing every 2 days policies, respectively. These correspond to increases of 180% (19.4 extra tests), 201% (32.9 extra tests) and 235% (64.4 extra tests) over no one in the population taking up the regular asymptomatic tests.

For the symptomatic experiments, we show the average number of tests per person taken over the 180 days of simulation in Figures G.8 and G.9.

Results are in agreement across different growth rates that more LFD tests are used on average per person. The exception lies when the growth rate is equal to 0.3, due to the significantly greater number of infections when using PCR tests, more tests of this type are required as more people get symptomatic and more people get contact traced. This increased requirement for PCR tests is more noticeable at high rates of adherence to testing after contact tracing: at 100% adherence to testing after contact tracing there are on average 66.3% more PCR tests consumed per person, whereas at 0% adherence there are on average 34.1% more LFD tests used per person.

Swapping PCR tests for LFD also impacts the amount of time each person spends in isolation, according to the results presented in Figure G.10 for experiments with

asymptomatic testing and Figure G.11 with symptomatic testing only.

In the asymptomatic case, we find that the difference between both tests is similar for 0.025, 0.05, 0.1 and 0.15 growth rates; people isolate longer on average when LFD tests are used, due to possibility of a false negative. In these growth rates, with 100% adherence to the testing policy, people spend on average 17.9% (0.6 days), 14.0% (0.5 days) and 7.8% (0.4 days) longer on isolation if LFD tests are used for the weekly, twice weekly and testing every 2 days policies, respectively. This difference is reduced, when the growth rate equals 0.225, to 2.9% (0.1 days), 2.4% (0.1 days) and 4.3% (0.2 days) and reversed when the growth rate is 0.3, leading to people isolating longer when PCR tests are used by 6.6% (0.5 days), 5.9% (0.4 days) and 5.9% (0.4 days). For the 0.3 growth rate, the longer isolation times when using PCR tests are caused by a greater number of infections, with the difference being large enough to be reflected in the number of person-days of isolation.

At higher levels of adherence to regular asymptomatic testing, the increase in the number of person-days of isolation by using a LFD test instead of PCR is smaller than at low levels of adherence. For example, for the twice weekly testing policy, in 0.025, 0.05, 0.1 and 0.15 growth rates, the average increase in isolation days when changing the PCR tests for LFD is 21.7% (0.6 days) at 0% adherence and 17.4% (0.6 days) at 50% adherence. This is a consequence of the greater number of false positives due to intensive asymptomatic testing at high adherence, with the majority of isolation time being caused by asymptomatic testing so changing to LFD testing in symptomatic (ILI or COVID) and contact traced individuals is less impactful in this metric.

Results also show that with LFD tests for symptomatic or contact traced individuals, the proportion of time spent isolating while infected with COVID is smaller across all growth rates and testing strategies. Using the same example as before, at a growth rate of 0.3 with POLYMOD contact patterns, for the testing every two days policy, the time spent isolating while positive varies from 49% at 0% adherence to 31% at 100%, and for the twice weekly testing policy, these values change to 51% and 44%, respectively. This is likely to be caused by a combination of better infection control (less infections leads to less time isolating while positive) and more false positives (which leads to more time isolating while negative for COVID).

A similar behaviour is observed in the symptomatic experiments, where the difference in number of person-days in isolation between testing with PCR and LFD is similar for 0.025, 0.05, 0.1 and 0.15 growth rates. We observe that contact tracing only affects the number of person-days of isolation when the growth rate equals 0.3,

where increasing adherence to testing after contact tracing using PCR tests leads to an average of 0.8 extra person-days of isolation, against 0.2 days when using LFD tests instead. This contrasts with the remaining growth rates, where increasing adherence leads to less than 0.1 extra person-days of isolation.

Results also show that increasing adherence to isolation on symptoms increases the difference in person-days of isolation between using PCR or LFD tests, as adherence to this intervention increases, using LFD tests leads to more person-days of isolation than when using PCR tests. For the 0.025, 0.05, 0.1 and 0.15 growth rates, people spend on average of 0.1 extra days in isolation when using LFD instead of PCR when adherence to isolation on symptoms is 0%, growing to 1.4 days when adherence is 100%. For the 0.225 and 0.3 growth rates, at low levels of adherence to isolation on symptoms, people spend more time on isolation when PCR tests are used, 0.5 days at 0.225 growth rate and 1.5 days. As adherence to isolation on symptoms increases, this value shifts so that using LFD tests leads to people spending more time on isolation, 1.0 extra days at 0.225 growth rate when adherence is 100% and 0.4 days at 0.3 growth rate.

Over all, the greater the adherence to isolation on symptom onset, the less time spent isolating while negative to COVID. This highlights how increasing adherence to isolation on symptom onset achieves a high reduction in number of infections: by preemptively isolating individuals with any symptoms that could be related to COVID, many susceptible people are being removed from the contact networks and, therefore, less likely to be infected.

Similarly to the asymptomatic experiments, when using LFD tests individuals spend more time isolating while they are negative to COVID. Excluding when there is no isolation on symptom onset, when using PCR tests individuals spent on average 5% of their time on isolation positive to COVID, with a growth of 0.025, up to 56%, with a growth rate of 0.3. When using LFD tests instead, these values are reduced to 3 and 47%, respectively.

5.4 Discussion

We looked at three different policies for asymptomatic testing, one based on UK government advice in Autumn 2021 that recommended LFD testing twice per week and two alternatives that doubled and halved the frequency of testing compared to this recommendation. These asymptomatic testing policies are compared against a scenario

of partial closure of schools and work places and reductions in community contacts, which could be implemented through a dissuasion of activities that require indoor mixing or placing limits on groups sizes in restaurants/pubs. These scenarios emulate lockdown scenarios at different levels and are naturally very effective at infection control. However, they carry a high cost in the form of economic recession and mental health impact on citizens that we do not account for in the simulations and it is unrealistic to consider that these measures of infection control could be maintained for as long as 6 months (the length of our simulations).

Even with an experimental setup that favours a reduction of contacts, we find that with the twice weekly testing schedule, increasing the amount of people participating in the regular asymptomatic testing can have as much impact as reducing the number of contacts by 30%. When halving/doubling the testing frequency this impact is reduced/increased to up to 20% or 45%, respectively.

The asymptomatic testing strategies carry a cost that can be measured by the number of tests each person takes and by how long each person spends in isolation. Perfect adherence to the twice weekly asymptomatic testing advice requires that on average each person in the whole population take approximately 50 tests over a 180 days period, the majority of which are LFD, and spend approximately 3.5 days isolating while not infected with COVID-19. These days spent in isolation are, for the most part, a consequence of preemptive isolation while waiting for a test result, due to developing symptoms (COVID-19 or ILI). The increase in testing, which identifies more infections but also carries a risk of generating more false positives, is only responsible for an increase of less than 1 person day of isolation.

We also explored the impact of different levels of adherence to infection control policies in a scenario where there is no asymptomatic testing in the population. In this context, we find that the most effective policy in decreasing the number of infections is preemptive isolation on symptom onset, while waiting for a test result. This highlights the need for widespread education of which are the most common symptoms of COVID, particularly if new variants emerge that show different profiles of symptoms, and government's support to facilitate this isolation period. It is also necessary to support high levels of testing to prevent long periods of isolation from people with ILI symptoms that overlap with COVID.

However, in a time with little or no appetite to support preemptive isolation periods, our experimental setup shows that high rates of symptomatic testing and adherence to testing after contact tracing can achieve a similar reduction in the number of infections

in the population. In fact, our assumptions of the percentage of contacts found through contact tracing is conservative according to estimates from UK NHS Test and Trace [55], so the effectiveness of the combination of symptomatic testing and contact tracing could be higher than the results from our simulations, especially considering that achieving higher adherences to these policies is easier than high adherence to preemptive isolation.

Finally, we studied the effect of replacing PCR tests for LFD in symptomatic and contact traced individuals and results show that in the majority of scenarios, this substitution leads to a smaller number of total infections. This difference is caused by the testing delay associated with PCR tests, as the need to develop the test results in a laboratory setting offsets the gain obtained by being more accurate at identifying positive individuals. The greater sensitivity of PCR tests is even less important when we factor in the ability to perform multiple LFD tests before learning the result of a PCR test, especially because the majority of tests done in this context are due to symptomatic individuals (more so than contact traced ones) so it becomes imperative to diagnose and isolate these individuals when they are at the peak of their infectivity.

The reduction in number of infections by replacing PCR tests by LFD is achieved by a trade-off that can be quantified through an increase in both the number of tests done and the average number of person days of isolation; the latter caused by false positive tests, leading to people spending a greater portion of their isolation periods when not infected with COVID. The increase in number of tests when using LFD instead of PCR is only noticeable in the experiments with no asymptomatic testing and is a consequence of the lower test sensitivity; individuals are more likely to test negative with a LFD test than with a PCR but, because they are symptomatic, they are likely to take more tests during their symptomatic period. This increase in number of tests is not enough to offset to cost disparity between PCR and LFD tests, as each LFD test is much cheaper than a PCR test and requires less infrastructure.

Our analysis has a number of strengths. Our results are similar for different contact patterns, implying that our conclusions do not require the assumption that average number of contacts stays as low as during the pandemic. We examine a range of growth rates to cover different scenarios of epidemic growth, allowing us to study the cases where no interventions are needed (growth rates of 0.025 and 0.05), when testing is most effective (growth rates of 0.1, 0.15 and 0.225) and when testing alone is not sufficient (growth rate of 0.3).

Our work has some limitations. Firstly, Covasim is a stochastic model and the

inherent stochasticity gives rise to uncertainty, which is compounded with uncertainty arising from the model parameters. For the purposes of this analysis we ran 100 seeds for each set of parameters, which gave enough robustness in generating central, low and high outcomes across the scenarios.

Secondly, in this study and when using Covasim in general, while most of the parameters are derived from the literature, there are gaps where we had made some assumptions. For example, for the asymptomatic experiments, we have assumed 40% background symptomatic testing, adherence to isolation on symptoms and adherence to testing after contact tracing. Although this assumption is based on the results of the symptomatic experiments, it is unclear the extent to which the amount of symptomatic testing affects the efficacy of asymptomatic testing. Another assumption in the asymptomatic experiments is how much asymptomatic testing people outside the regular testing cohort should perform. A sensitivity analysis on the chosen parameters (90/10% split for intermittent and anti testers) would be useful to understand their impact on our results.

Similarly, while we have made an effort to quantify vaccine efficacy against onward transmission of different variants and waning protection from both vaccination and infection, current data is not certain on these. Specifically, the version of Covasim used in this study assumed a single variant (albeit with different step-like change in transmissibility) and a single antibody waning function for all individuals and all types of immunity, with individual- and immune-level variation in the level of Neutralising Antibodies (NAbs). Incorporating different – possibly competing variants – and different NAbs trajectories was beyond the scope of this work.

5.5 Conclusion

In summary, we explored the impact of different testing strategies using the Covasim ABM. Our results highlight that asymptomatic testing with LFTs is particularly effective when the growth rate corresponds to a weekly doubling in number of cases. We also found that swapping PCR tests for LFD may help in epidemic control due to the delay in returning test results to individuals, which seems to offset the loss in test sensitivity in LFTs.

Overall our findings highlight that regular asymptomatic testing with LFTs can be an impactful intervention to reduce COVID-19 burden and as an alternative to national

lockdowns. Delivering a strategy of large scale testing with LFTs combined with immediate self-quarantining and contact tracing of contacts will be crucial in dampening transmission, reducing hospitalisations and deaths in any future COVID-19 outbreaks and as countries move towards “living with COVID-19” strategy.

Chapter 6

Conclusion

6.1 Summary of results

Throughout this thesis, we explored how graphlets, small units of functional behaviour in the network, are an useful topological component of complex networks for the task of network comparison. Advances in algorithms and computational resources have allowed the adoption of graphlets as a network analysis technique for large scale networks, but their usage has not yet been embraced by research domains that could benefit from the insights graphlets provide.

From a methodology perspective, we proposed an extension of the network comparison measure based on graphlets to directed networks. The measure, NetEmd, had been previous shown to be more accurate than other methods at separating networks generated from different random graph models, even when they differ greatly in size, as well as identifying real world networks from the same domain. The main challenge of this extension is the combinatorial explosion in number of orbits due to edge direction. One way to deal with this issue is to include in the comparison only the set of orbits that occur in either of the networks under comparison. An alternative method is to consider dimensionality reduction techniques, such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA), as de-noising procedures in a dataset where we do not have a good noise model.

We presented extensive experimental results that demonstrate the performance of the extensions we propose. In particular when using ICA as the de-noising method, results show significant improvements in the ability to identify networks generated from the same random graph model, even when the networks differ in size by an order of magnitude, for undirected networks and for different levels of reciprocity in directed

networks. Our experimental setup also serves as a guide to which parametrization of NetEmd performs best in each context and when the added computational load of larger graphlets can be avoided in favour of using orbits from smaller graphlets only.

From an application perspective, we studied how vaccination discussions are conducted in an online discussion board, emphasizing the difference between these discussions in the context of COVID-19 against vaccination discussions before the pandemic. We found that the structure of the network varies over time depending on real world events that affect these discussions. We also found evidence of correlation between differences in network structure and the sentiment towards vaccination. The link between these two different types of analysis can be a useful guide to more efficiently allocate resources when performing sentiment labeling, an expensive and labourious process.

Network comparison was also found to be useful when modeling the impact of Test-Trace-Isolate (TTI) strategies for COVID-19, in support of policy for the United Kingdom (UK) government. In this context, we implemented a generation mechanism for social contact networks informed by age-mixing matrices, which are derived from social contact surveys representative of the UK population that measure the average number of social contacts of each participant depending on the transmission setting. This network generation mechanism was included in Covasim, an agent based model with multiple network layers that represent common places where transmission of the virus can take place, namely households, work places, schools or general community areas. Through network comparison, we showed the different patterns of contacts that emerge from using age-mixing matrices as a generation mechanism and how the average number of contacts affects the structure of those networks.

Using Covasim, we investigated different strategies for asymptomatic testing and adherence to Test-Trace-Isolate (TTI) interventions, such as symptomatic testing, isolation on symptoms or uptake of a test after contact tracing, measuring their impact on COVID-19 outcomes. We found that intensive asymptomatic testing can have as much impact as a lockdown scenario where people are asked to reduce their non-household contacts by 45%. Another result from this analysis was the greater ability of Lateral Flow Tests (LFTs) to control transmission of COVID-19 when compared to Polymerase Chain Reaction (PCR) tests, due to the delay in obtaining the test result when using the latter.

6.2 Discussion

The contributions of this thesis are split into three distinct areas, as delineated by the research questions in the Introduction. These three areas can generally be described as: methods to compare complex networks based on graphlet distributions, applications of network comparison in social networks and agent-based simulation of COVID-19 epidemics.

In the methodology chapter, although we have shown how our proposed algorithms improved upon the state of the art, it is remarkable how accurate previous methods are. Particularly in differentiating random network models, our experiments indicate that all the methods we analyze are able to distinguish correctly different random network models when they are generated with the same number of nodes and edges. This shows that graphlets are a powerful tool in characterizing networks and descriptive of their organizational structure. But it also shows that the framework proposed by Yaveroğlu et al. [193] has become more of a proof of concept rather than a proper way to evaluate network comparison methods, highlighting the need to develop benchmarking datasets for this task, as is common in graph learning or community detection tasks.

Still regarding methodology of network comparison, one important byproduct of our experimental setup, in the systematic manner we acquire results using orbits from multiple graphlet sizes, is the ability to tell whether scaling up to larger graphlets is worth the added computational load. A surprising result, that matches the findings of Wegner et al. [180], is that smaller graphlets are better at telling apart networks that differ greatly in size. A possible reason for this is that the networks we compare in our experimental analysis may be too small to obtain enough realizations from the distribution of these larger graphlets, so the approximations we observe in the network we generated are not in enough to fully characterize the distributions. This is also a reason why we observe a large gain in performance from the denoising methods we propose, by removing some of the variability introduced by the generation mechanism it becomes easier to distinguish different distributions and identify similar ones.

In the application chapter, we must emphasize the connection we found between differences in network structure to changes in sentiment towards vaccination, connecting two distinct techniques of analyzing discussions in social networks. Much of the work in analyzing vaccination discussions focuses in either the structure of discussions or their content but, by showing that these two components are connected, our work suggests that the two types of analysis can be used to complement each other. By

looking at differences in how users are connected in these discussions, we can generate nuanced hypothesis as to why the content or sentiment of their conversations has shifted. Conversely, by noting changes in sentiment, we can obtain context to the differences in distributions of re-occurring patterns of connections between users. In sum, our analysis of vaccination discussions within Mumsnet shows how network comparison based on graphlets can be used as a tool to generate hypothesis about changes in behaviour of users in a social network.

Our analysis of vaccination discussions in Mumsnet, split between pre and post COVID-19, also shows the impact of the global pandemic on vaccination discourse online. Whereas discussions about vaccination before COVID-19 were a niche topic in Mumsnet, with a small but loyal subset of users that engaged in these discussions over multiple years, discussions about COVID-19 vaccinations were more popular, with users maintaining longer periods of engagement with the forum. This is likely due to how disruptive the pandemic was worldwide, increasing the willingness to discuss the topic online. Another key difference between the periods before and after COVID-19 is the volatility of sentiment towards vaccination, as before COVID-19 we observed time periods where discussion was dominated by either negative or positive sentiments towards vaccination, but after COVID-19, the most common sentiment was neutrality, possibly due to our wide criteria for inclusion of discussions.

A limitation of our study was the inability to find a connection between specific orbits or graphlets and prevalence of a particular sentiment. Although we were able to detect differences in the structure of the user interaction networks, often these differences were tied to how the graphlets were spread over the network, which impacts the orbit distributions in subtle ways that makes the hypothesis generation process more difficult.

Finally, on the agent-based simulation of COVID-19, although we did not explore in depth the impact of including social contact surveys like CoMix and POLYMOD to inform the contact networks of the agents, the calibration of parameters epidemic parameters suggests that the reduced number of contacts observed during the COVID-19 pandemic requires a much higher rate of transmission between individuals to reach the same final epidemic size. Although this finding seems obvious, it is important for models to reflect reality when exploring strategies for mitigation and the impact of possible policy.

The research questions about simulation of outcomes for COVID-19 in this thesis were informed by policy makers in the UK government and posed to us by the scientific

committees that inform these policy makers. Thus, the focus of the final research chapter of this thesis is more centered on the effects of different testing and isolation strategies on various outcomes, than on the impact of network structure in Covasim. In this context, our findings highlight that background asymptomatic testing with Lateral Flow Device (LFD) testing is a viable strategy to contain an epidemic as an alternative for mass lockdowns, for which the public currently has little appetite for. It is also clear from our results that the advantages of LFTs over PCR tests (immediate test results and ease of repeated testing) overcome the lower test sensitivity in the ability to control a COVID-19 epidemic.

6.3 Future work

The work we presented in each chapter has the potential to seed future lines of research, some of which we propose below:

- When denoising orbit frequencies for NetEmd, we explored two linear models of dimensionality reduction, PCA and ICA. However, graphs are notably a non-linear data structure, so non-linearity is to be expected when considering orbit frequencies. In particular, future work could explore using autoencoders [85] for the denoising process, as they have been proposed as generalizations of PCA for non-linear data.

- In Chapter 4, we mention the potential of combining the information contained in which orbits are driving changes in the network structure with how the sentiment is changing in the forum to generate hypothesis about user behaviour. Future work should approach this potential from a multi-disciplinary perspective, encompassing the feedback from sociology experts to generate those hypothesis.

- The configuration model used to generate the network for Covasim is an implementation that does not guarantee known properties of real world networks, such as scale-free [3] or small-world [179]. Both these properties potentially affect how a virus can spread in a social contact network, for instance, in a scale-free network, there is a small group of people that accrues a disproportionate number of connections. If one person in this group is infected, then they have a large neighbourhood of people they can potentially spread the virus to, a “super spreader" event. Future work should implement random graph models that

better reflect how a social contact network grows, for example, a preferential attachment model [9].

Bibliography

[1] Aseel Addawood. 2018. Usage of scientific references in MMR vaccination debates on Twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 971–979.

[2] Tania M Alarcon Falconi, Bertha Estrella, Fernando Sempértegui, and Elena N Naumova. 2020. Effects of data aggregation on time series analysis of seasonal infections. *International journal of environmental research and public health* 17, 16 (2020), 5887.

[3] Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics* 74, 1 (2002), 47.

[4] Waqar Ali, Tiago Rito, Gesine Reinert, Fengzhu Sun, and Charlotte M Deane. 2014. Alignment-free protein interaction network comparison. *Bioinformatics* 30, 17 (2014), i430–i437.

[5] Waqar Ali, Anatol E Wegner, Robert E Gaunt, Charlotte M Deane, and Gesine Reinert. 2016. Comparison of large networks with sub-sampling strategies. *Scientific Reports* 6, 1 (2016), 1–8.

[6] David Aparício, Pedro Ribeiro, Tijana Milenković, and Fernando Silva. 2019. Temporal network alignment via GoT-WAVE. *Bioinformatics* 35, 18 (02 2019), 3527–3529.

[7] David Aparício, Pedro Ribeiro, and Fernando Silva. 2016. Extending the applicability of graphlets to directed networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 14, 6 (2016), 1302–1315.

[8] David Aparício, Pedro Ribeiro, and Fernando Silva. 2018. Graphlet-orbit Transitions (GoT): A fingerprint for temporal network comparison. *PloS One* 13, 10 (2018), e0205497.

[9] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science* 286, 5439 (1999), 509–512.

[10] Christopher L Barrett, Keith R Bisset, Stephen G Eubank, Xizhou Feng, and Madhav V Marathe. 2008. Episimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *SC'08: Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*. IEEE, 1–12.

[11] Gema Bello-Orgaz, Julio Hernandez-Castro, and David Camacho. 2017. Detecting discussion communities on vaccination in twitter. *Future Generation Computer Systems* 66 (2017), 125–136.

[12] Elizabeth B Blankenship, Mary Elizabeth Goff, Jinging Yin, Zion Tsz Ho Tse, King-Wa Fu, Hai Liang, Nitin Saroha, and Isaac Chun-Hai Fung. 2018. Sentiment, contents, and retweets: A study of two vaccine-related twitter datasets. *The Permanente Journal* 22 (2018).

[13] Stuart Blume. 2006. Anti-vaccination movements and their interpretations. *Social Science & Medicine* 62, 3 (2006), 628–642.

[14] Stephen P Borgatti and Martin G Everett. 1997. Network analysis of 2-mode data. *Social Networks* 19, 3 (1997), 243–270.

[15] Alan P Boyle, Carlos L Araya, Cathleen Brdlik, Philip Cayting, Chao Cheng, Yong Cheng, Kathryn Gardner, LaDeana W Hillier, Judith Janette, Lixia Jiang, et al. 2014. Comparative analysis of regulatory information and circuits across distant species. *Nature* 512, 7515 (2014), 453–456.

[16] Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32.

[17] Emily K Brunson. 2013. The impact of social networks on parents' vaccination decisions. *Pediatrics* 131, 5 (2013), e1397–e1404.

[18] Peter Buckle, Massimo Micocci, John Tulloch, Patrick Kierkegaard, Paula Parvulescu, Carl Thompson, Karen Spilsbury, oy Allen, Richard Body, Gail Hayward, Iain Buchan, and Adam L Gordon. 2021. COVID-19 point-of-care testing in care homes: what are the lessons for policy and practice? *Age and Aging* 50 (2021). Issue 5.

[19] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2016. *Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data?* American Psychological Association, 133—-139.

[20] Helen Campbell, Angela Edwards, Louise Letley, Helen Bedford, Mary Ramsay, and Joanne Yarwood. 2017. Changing attitudes to childhood immunisation in english parents. *Vaccine* 35, 22 (2017), 2979–2985.

[21] Wei-Chu Chen and Staša Milojevic. 2018. Interaction or segregation: Vaccination and information sharing on twitter. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 301–304.

[22] Sarvenaz Choobdar, Pedro Ribeiro, Sylwia Bugla, and Fernando Silva. 2012. Comparison of co-authorship networks across scientific fields using motifs. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE, 147–152.

[23] Jamie Cohen, Robyn Stuart, Katherine Rosenfeld, Hil Lyons, Michael White, Cliff Kerr, Daniel Klein, and Michael Famulare. 2021. Quantifying the role of naturally-and vaccine-derived neutralizing antibodies as a correlate of protection against COVID-19 variants. *medRxiv* (2021). https://www.medrxiv.org/content/early/2021/09/28/2021.05.31.21258018

[24] The OpenSAFELY Collaborative, Helen J Curtis, and Peter et al. Inglesby. 2021. Trends and clinical characteristics of COVID-19 vaccine recipients: a federated analysis of 57.9 million patients' primary care records in situ using OpenSAFELY. *medRxiv* (2021). https://doi.org/10.1101/2021.01.25.21250356

[25] Pierre Comon. 1994. Independent component analysis, a new concept? *Signal Processing* 36, 3 (1994), 287–314.

[26] Stephen A. Cook. 1971. The complexity of theorem-proving procedures. In *Proceedings of the Third Annual ACM Symposium on Theory of Computing*. 151–158.

[27] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20, 3 (1995), 273–297.

[28] Eleonora D'Andrea, Pietro Ducange, Alessio Bechini, Alessandro Renda, and Francesco Marcelloni. 2019. Monitoring the public opinion about the vaccination

topic from tweets analysis. *Expert Systems with Applications* 116 (2019), 209–226.

[29] Paul Davies, Simon Chapman, and Julie Leask. 2002. Antivaccination activists on the world wide web. *Archives of Disease in Childhood* 87, 1 (2002), 22–25.

[30] Emma L Davis, Tim C D Lucas, Anna Borlase, Timothy M Pollington, Sam Abbott, Diepreye Ayabina, Thomas Crellen, Joel Hellewell, Li Pi, CMMID COVID-19 Working Group, Graham F Medley, T Deirdre Hollingsworth, and Petra Klepac. 2021. Contact tracing is an imperfect tool for controlling COVID-19 transmission and relies on population adherence. *Nature Communications* 12 (2021). Issue 5412.

[31] Michael S Deiner, Cherie Fathy, Jessica Kim, Katherine Niemeyer, David Ramirez, Sarah F Ackley, Fengchen Liu, Thomas M Lietman, and Travis C Porco. 2019. Facebook and Twitter vaccine sentiment in response to measles outbreaks. *Health informatics journal* 25, 3 (2019), 1116–1132.

[32] Federico Delussu, Michele Tizzoni, and Laetitia Gauvin. 2022. Evidence of pandemic fatigue associated with stricter tiered COVID-19 restrictions. *PLOS Digital Health* 1, 5 (2022), e0000035.

[33] Sarah Denford, Fiona Mowbray, Lauren Towler, Helena Wehling, Gemma Lasseter, Richard Amlôt, Isabel Oliver, Lucy Yardley, and Matthew Hickman. 2022. Exploration of attitudes regarding uptake of COVID-19 vaccines among vaccine hesitant adults in the UK: a qualitative analysis. *BMC Infectious Diseases* 22, 1 (2022), 1–14.

[34] Sarah Schaffer DeRoo, Natalie J Pudalov, and Linda Y Fu. 2020. Planning for a COVID-19 vaccination program. *JAMA* 323, 24 (2020), 2458–2459.

[35] Catherine Dibble, Stephen Wendel, and Kristofor Carle. 2007. Simulating pandemic influenza risks of US cities. In *2007 Winter Simulation Conference*. IEEE, 1548–1550.

[36] Jingcheng Du, Jun Xu, Hsingyi Song, Xiangyu Liu, and Cui Tao. 2017. Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. *Journal of Biomedical Semantics* 8, 1 (2017), 1–7.

[37] Louise Dyson, Edward M. Hill, Sam Moore, Jacob Curran-Sebastian, Michael J. Tildesley, Katrina A. Lythgoe, Thomas House, Lorenzo Pellis, and Matt J. Keeling. 2021. Possible future waves of SARS-CoV-2 infection generated by variants of concern with a range of characteristics. *Nature Communications* 12, 1 (2021), 5730.

[38] Paul Erdős and Alfréd Rényi. 1960. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5, 1 (1960), 17–60.

[39] Luca Ferretti, Alice Ledda, Chris Wymant, Lele Zhao, Virginia Ledda, Lucie Abeler-Dörner, Michelle Kendall, Anel Nurtay, Hao-Yuan Cheng, Ta-Chou Ng, Hsien-Ho Lin, Rob Hinch, Joanna Masel, A. Marm Kilpatrick, and Christophe Fraser. 2020. The timing of COVID-19 transmission. *medRxiv* (2020). `https://doi.org/10.1101/2020.09.04.20188516`

[40] Luca Ferretti, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abele-Dorner, Michael Parker, David Bonsall, and Christophe Fraser. 2020. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* 368 (2020). Issue 6491.

[41] Jason Flannick, Antal Novak, Chuong B Do, Balaji S Srinivasan, and Serafim Batzoglou. 2008. Automatic parameter learning for multiple network alignment. In *Annual International Conference on Research in Computational Molecular Biology*. Springer, 214–231.

[42] Abigail J Ford and Nisreen A Alwan. 2018. Use of social networking sites and women's decision to receive vaccinations during pregnancy: A cross-sectional study in the UK. *Vaccine* 36, 35 (2018), 5294–5303.

[43] Santo Fortunato. 2010. Community detection in graphs. *Physics Reports* 486, 3-5 (2010), 75–174.

[44] Linton C Freeman. 1978. Centrality in social networks conceptual clarification. *Social Networks* 1, 3 (1978), 215–239.

[45] Sebastian Funk, Sam Abbott, Benjamin D Atkins, Marc Baguelin, J Kenneth Baillie, P Birrell, Joshua Blake, Nikos I Bosse, Joshua Burton, Jonathan Carruthers, et al. 2020. Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. *MedRxiv* (2020).

[46] Martyn Fyles, Elizabeth Fearon, Christopher Overton, University of Manchester COVID-19 Modelling Group, Tom Wingfield, Graham F Medley, Ian Hall, Lorenzo Pellis, and Thomas House. 2021. Using a household-structured branching process to analyse contact tracing in the SARS-CoV-2 pandemic. *Royal Society Philosophical Transactions B* 376 (2021). Issue 1829.

[47] Wilfredo F. Garcia-Beltran, Kerri J. St. Denis, and Angelique Hoelzemer et al. 2022. mRNA-based COVID-19 vaccine boosters induce neutralizing immunity against SARS-CoV-2 Omicron variant. *Cell* 185, 3 (2022), 457–466.e4. https://www.sciencedirect.com/science/article/pii/S0092867421014963

[48] Edward N Gilbert. 1961. Random plane networks. *J. Soc. Indust. Appl. Math.* 9, 4 (1961), 533–543.

[49] Daniel T Gillespie. 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics* 22, 4 (1976), 403–434.

[50] Daniel T Gillespie. 1977. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry* 81, 25 (1977), 2340–2361.

[51] Amy Gimma, James D. Munday, Kerry L. M. Wong, Pietro Coletti, Kevin van Zandvoort, Kiesha Prem, CMMID COVID-19 working group, Petra Klepac, G. James Rubin, Sebastian Funk, W. John Edmunds, and Christopher I. Jarvis. 2022. Changes in social contacts in England during the COVID-19 pandemic between March 2020 and March 2021 as measured by the CoMix survey: A repeated cross-sectional study. *PLOS Medicine* 19, 3 (2022), e1003907.

[52] GOV.UK. 2021. Reproduction number (R) and growth rate: methodology. (2021). https://www.gov.uk/government/publications/reproduction-number-r-and-growth-rate-methodology/reproduction-number-r-and-growth-rate-methodology

[53] GOV.UK. 2022. Coronavirus (COVID-19) in the UK: UK Summary. (2022). https://coronavirus.data.gov.uk

[54] GOV.UK. 2022. Vaccinations in the UK. https://coronavirus.data.gov.uk/details/vaccinations. (2022). Accessed 03-05-2022.

[55] GOV.UK. 2022. Weekly statistics for NHS Test and Trace (England). (2022). https://www.gov.uk/government/collections/nhs-test-and-trace-statistics-england-weekly-reports

[56] John J Grefenstette, Shawn T Brown, Roni Rosenfeld, Jay DePasse, Nathan TB Stone, Phillip C Cooley, William D Wheaton, Alona Fyshe, David D Galloway, Anuroop Sriram, et al. 2013. FRED (A Framework for Reconstructing Epidemic Dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC public health* 13, 1 (2013), 1–14.

[57] Shawn Gu, John Johnson, Fazle E Faisal, and Tijana Milenković. 2018. From homogeneous to heterogeneous network alignment via colored graphlets. *Scientific Reports* 8, 1 (2018), 1–16.

[58] Douglas Guilbeault, Joshua Becker, and Damon Centola. 2018. *Complex contagions: a decade in review*. Springer International Publishing, Cham, Switzerland, 3–25.

[59] Wayne Hayes, Kai Sun, and Nataša Pržulj. 2013. Graphlet-based measures are suitable for biological network comparison. *Bioinformatics* 29, 4 (2013), 483–491.

[60] Joel Hellewell, Timothy W Russell, Rupert Beale, Gavin Kelly, Catherine Houlihan, Eleni Nastouli, and Adam J Kucharski. 2021. Estimating the effectiveness of routine asymptomatic PCR testing at different frequencies for the detection of SARS-CoV-2 infections. *BMC medicine* 19 (2021), 106.

[61] Joel Hellewell, Timothy W Russell, Catherine Houlihan, Eleni Nastouli, and Adam J Kucharski. 2020. Estimating effectiveness of frequent PCR testing at different intervals for the detection of SARS-CoV-2 infections. (2020). https://cmmid.github.io/topics/covid19/pcr-positivity-over-time.html

[62] Desmond J Higham, Marija Rašajski, and Nataša Pržulj. 2008. Fitting a geometric graph to a protein–protein interaction network. *Bioinformatics* 24, 8 (2008), 1093–1099.

[63] Robert Hinch, William JM Probert, Anel Nurtay, Michelle Kendall, Chris

Wymant, Matthew Hall, Katrina Lythgoe, Ana Bulas Cruz, Lele Zhao, Andrea Stewart, et al. 2021. OpenABM-Covid19—An agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. *PLoS computational biology* 17, 7 (2021), e1009146.

[64] G. E. Hinton and R. R. Salakhutdinov. 2006. Reducing the Dimensionality of Data with Neural Networks. *Science* 313, 5786 (2006), 504–507.

[65] Tomaž Hočevar and Janez Demšar. 2014. A combinatorial approach to graphlet counting. *Bioinformatics* 30, 4 (2014), 559–565.

[66] Beth L. Hoffman, Elizabeth M. Felter, Kar-Hai Chu, Ariel Shensa, Chad Hermann, Todd Wolynn, Daria Williams, and Brian A. Primack. 2019. It's not all about autism: The emerging landscape of anti-vaccination sentiment on Facebook. *Vaccine* 37, 16 (2019), 2216–2223. https://www.sciencedirect.com/science/article/pii/S0264410X19303032

[67] Brian S Hooker. 2014. Measles-mumps-rubella vaccination timing and autism among young african american boys: a reanalysis of CDC data. *Translational Neurodegeneration* 3, 1 (2014), 1–6.

[68] Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24, 6 (1933), 417.

[69] Thomas House, Geoffrey Davies, Leon Danon, and Matt J Keeling. 2009. A motif-based approach to network epidemics. *Bulletin of Mathematical Biology* 71, 7 (2009), 1693—1706.

[70] Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification* 2, 1 (1985), 193–218.

[71] Clayton Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8. 216–225.

[72] Aapo Hyvärinen. 1999. The fixed-point algorithm and maximum likelihood estimation for independent component analysis. *Neural Processing Letters* 10, 1 (1999), 1–5.

[73] Aapo Hyvärinen and Erkki Oja. 2000. Independent component analysis: algorithms and applications. *Neural Networks* 13, 4-5 (2000), 411–430.

[74] Iaroslav Ispolatov, Pavel L Krapivsky, and Anton Yuryev. 2005. Duplication-divergence model of protein interaction network. *Physical Review E* 71, 6 (2005), 061911.

[75] Takashi Ito, Tomoko Chiba, Ritsuko Ozawa, Mikio Yoshida, Masahira Hattori, and Yoshiyuki Sakaki. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences* 98, 8 (2001), 4569–4574.

[76] S Mo Jang, Brooke W Mckeever, Robert Mckeever, and Joon Kyoung Kim. 2019. From social media to mainstream news: the information flow of the vaccine-autism controversy in the US, Canada, and the UK. *Health Communication* 34, 1 (2019), 110–117.

[77] C Jarvis, J Munday, A Gimma, K Wong, K Van Zandvoort, S Funk, et al. 2021. Social contacts in the UK from the CoMix social contact survey: Report for survey week 61. *London School of Tropical Medicine and Hygiene* 1 (2021).

[78] Christopher I Jarvis, Kevin Van Zandvoort, Amy Gimma, Kiesha Prem, Petra Klepac, G James Rubin, and W John Edmunds. 2020. Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK. *BMC medicine* 18 (2020), 124.

[79] Ian T Jolliffe. 2002. *Principal component analysis*. Springer-Verlag.

[80] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 427–431.

[81] Ruian Ke, Pamela P. Martinez, Rebecca L. Smith, Laura L. Gibson, Agha Mirza, Madison Conte, Nicholas Gallagher, Chun Huai Luo, Junko Jarrett, Ruifeng Zhou, Abigail Conte, Tongyu Liu, Mireille Farjo, Kimberly K. O. Walden, Gloria Rendon, Christopher J. Fields, Leyi Wang, Richard Fredrickson, Darci C. Edmonson, Melinda E. Baughman, Karen K. Chiu, Hannah Choi, Kevin R. Scardina, Shannon Bradley, Stacy L. Gloss, Crystal Reinhart, Jagadeesh Yedetore,

Jessica Quicksall, Alyssa N. Owens, John Broach, Bruce Barton, Peter Lazar, William J. Heetderks, Matthew L. Robinson, Heba H. Mostafa, Yukari C. Manabe, Andrew Pekosz, David D. McManus, and Christopher B. Brooke. 2022. Daily longitudinal sampling of SARS-CoV-2 infection reveals substantial heterogeneity in infectiousness. *Nature Microbiology* 7, 5 (2022), 640–652.

[82] Cliff C. Kerr, Robyn M. Stuart, Dina Mistry, Romesh G. Abeysuriya, Katherine Rosenfeld, Gregory R. Hart, Rafael C. Núñez, Jamie A. Cohen, Prashanth Selvaraj, Brittany Hagedorn, Lauren George, Michał Jastrzębski, Amanda S. Izzo, Greer Fowler, Anna Palmer, Dominic Delport, Nick Scott, Sherrie L. Kelly, Caroline S. Bennette, Bradley G. Wagner, Stewart T. Chang, Assaf P. Oron, Edward A. Wenger, Jasmina Panovska-Griffiths, Michael Famulare, and Daniel J. Klein. 2021. Covasim: An agent-based model of COVID-19 dynamics and interventions. *PLOS Computational Biology* 17, 7 (2021), e1009149.

[83] István Z Kiss, Joel C Miller, Péter L Simon, et al. 2017. Mathematics of epidemics on networks. *Cham: Springer* 598 (2017), 31.

[84] Danai Koutra, Joshua T Vogelstein, and Christos Faloutsos. 2013. Deltacon: A principled massive-graph similarity function. In *Proceedings of the 2013 SIAM International Conference on Data Mining*. 162–170.

[85] Mark A Kramer. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal* 37, 2 (1991), 233–243.

[86] Oleksii Kuchaiev, Tijana Milenković, Vesna Memišević, Wayne Hayes, and Nataša Pržulj. 2010. Topological network alignment uncovers biological function and phylogeny. *Journal of the Royal Society Interface* 7, 50 (2010), 1341–1354.

[87] Oleksii Kuchaiev and NATAŠA PRŽULJ. 2009. Learning the structure of protein-protein interaction networks. In *Biocomputing 2009*. World Scientific, 39–50.

[88] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th International Conference on World Wide Web*. 591–600.

[89] Daniel B. Larremore, Bryan Wilder, Evan Lester, Soraya Shehata, James M. Burke, James A. Hay, Milind Tambe, Michael J. Mina, and Roy Parker. 2021.

Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. *Science Advances* 7, 1 (2021), eabd5393.

[90] Trystan Leng, Edward M Hill, Robin N Thompson, Michael J Tildesley, Matt J. Keeling, and Louise Dyson. 2022. Assessing the impact of lateral flow testing strategies on within-school SARS-CoV-2 transmission and absences: A modelling study. *PLoS Computational Biology* 18 (2022). Issue 5.

[91] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2005. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. 177–187.

[92] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2007. Graph evolution: densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 2–es.

[93] Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. `http://snap.stanford.edu/data`. (June 2014).

[94] Lun Li, David Alderson, John C Doyle, and Walter Willinger. 2005. Towards a theory of scale-free graphs: definition, properties, and implications. *Internet Mathematics* 2, 4 (2005), 431–523.

[95] Zhongjie Li, Fengfeng Liu, Jinzhao Cui, Zhibin Peng, Zhaorui Chang, Shengjie Lai, Qiulan Chen, Liping Wang, George F. Gao, and Zijian Feng. 2021. Comprehensive large-scale nucleic acid–testing strategies support China's sustained containment of COVID-19. *Nature Medicine* 27, 5 (2021), 740–742.

[96] Fengchen Liu, Wayne TA Enanoria, Jennifer Zipprich, Seth Blumberg, Kathleen Harriman, Sarah F Ackley, William D Wheaton, Justine L Allpress, and Travis C Porco. 2015. The role of vaccination coverage, individual behaviors, and the public health response in the control of measles epidemics: an agent-based simulation for California. *BMC public health* 15, 1 (2015), 1–16.

[97] Liverpool Covid-SMART Community Testing Pilot. 2021. Evaluation Report 17 June. (2021). `https://www.liverpool.ac.uk/media/livacuk/research/Mass,testing,evaluation.pdf`

[98] Brad Love, Itai Himelboim, Avery Holton, and Kristin Stewart. 2013. Twitter as a source of vaccination information: content drivers and what they are saying. *American Journal of Infection Control* 41, 6 (2013), 568–570.

[99] Charles M Macal and Michael J North. 2005. Tutorial on agent-based modeling and simulation. In *Proceedings of the Winter Simulation Conference, 2005.* IEEE, 14–pp.

[100] Owen Macindoe and Whitman Richards. 2010. Graph comparison using fine structure analysis. In *2010 IEEE Second International Conference on Social Computing*. IEEE, 193–200.

[101] Nil Mamano and Wayne B Hayes. 2017. SANA: simulated annealing far outperforms many other search algorithms for biological network alignment. *Bioinformatics* 33, 14 (2017), 2156–2164.

[102] Shmoolik Mangan and Uri Alon. 2003. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences* 100, 21 (2003), 11980–11985.

[103] GC Marshall, R Skeva, C Jay, MEP Silva, M Fyles, T House, EL Davis, L Pi, GF Medley, BJ Quilty, L Dyson, L Yardley, and E Fearon. 2022. Public perceptions and interactions with UK COVID-19 Test, Trace and Isolate policies, and implications for pandemic infectious disease modelling [version 1; peer review: awaiting peer review]. *F1000Research* 11, 1005 (2022). `https://doi.org/10.12688/f1000research.124627.1`

[104] Alberto JM Martin, Calixto Dominguez, Sebastián Contreras-Riquelme, David S Holmes, and Tomas Perez-Acle. 2016. Graphlet based metrics for the comparison of gene regulatory networks. *PloS One* 11, 10 (2016), e0163497.

[105] Julian J McAuley and Jure Leskovec. 2012. Learning to discover social circles in ego networks. *Advances in Neural Information Processing Systems* 2012 (2012), 548–556.

[106] Cassie McMillan and Diane Felmlee. 2020. Beyond dyads and triads: a comparison of tetrads in twenty social networks. *Social Psychology Quarterly* 83, 4 (2020), 383–404.

[107] Tijana Milenković, Weng Leong Ng, Wayne Hayes, and Nataša Pržulj. 2010. Optimal network alignment with graphlet degree vectors. *Cancer informatics* 9 (2010), CIN–S4744.

[108] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. 2004. Superfamilies of evolved and designed networks. *Science* 303, 5663 (2004), 1538–1542.

[109] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. 2002. Network motifs: simple building blocks of complex networks. *Science* 298, 5594 (2002), 824–827.

[110] Bjarke Mønsted and Sune Lehmann. 2019. Algorithmic detection and analysis of vaccine-denialist sentiment clusters in social networks. *arXiv preprint arXiv:1905.12908* (2019).

[111] Joël Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, et al. 2008. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS medicine* 5, 3 (2008), e74.

[112] Martin Müller and Marcel Salathé. 2020. Addressing machine learning concept drift reveals declining vaccine sentiment during the COVID-19 pandemic. *arXiv preprint arXiv:2012.02197* (2020).

[113] Martin M Müller and Marcel Salathé. 2019. Crowdbreaks: tracking health trends using public social media data and crowdsourcing. *Frontiers in Public Health* 7 (2019), 81.

[114] Mark Newman. 2018. *Networks*. Oxford University Press.

[115] Mark EJ Newman. 2001. Scientific collaboration networks. II. shortest paths, weighted networks, and centrality. *Physical Review E* 64, 1 (2001), 016132.

[116] Mark Ed Newman, Albert-László Ed Barabási, and Duncan J Watts. 2006. *The structure and dynamics of networks*. Princeton University Press.

[117] Saad B. Omer, Daniel A. Salmon, Walter A. Orenstein, M. Patricia deHart, and Neal Halsey. 2009. Vaccine refusal, mandatory immunization, and the risks of vaccine-preventable diseases. *New England Journal of Medicine* 360, 19 (2009), 1981–1988.

[118] Jukka-Pekka Onnela, Daniel J Fenn, Stephen Reid, Mason A Porter, Peter J Mucha, Mark D Fricker, and Nick S Jones. 2012. Taxonomies of networks from community structure. *Physical Review E* 86, 3 (2012), 036104.

[119] Kate K Orroth, Esther E Freeman, Roel Bakker, Anne Buvé, Judith R Glynn, Marie-Claude Boily, Richard G White, J Dik F Habbema, and Richard J Hayes. 2007. Understanding the differences between contrasting HIV epidemics in east and west Africa: results from a simulation model of the Four Cities Study. *Sexually transmitted infections* 83, suppl 1 (2007), i5–i16.

[120] Christopher E Overton, Helena B Stage, Shazaad Ahmad, Jacob Curran-Sebastian, Paul Dark, Rajenki Das, Elizabeth Fearon, Timothy Felton, Martyn Fyles, Nick Gent, et al. 2020. Using statistics and mathematical modelling to understand infectious disease outbreaks: COVID-19 as an example. *Infectious Disease Modelling* 5 (2020), 409–441.

[121] Giuliano Andrea Pagani and Marco Aiello. 2013. The power grid as a complex network: a survey. *Physica A: Statistical Mechanics and its Applications* 392, 11 (2013), 2688–2700.

[122] Jasmina Panovska-Griffiths, Cliff C Kerr, Robyn M Stuart, Dina Mistry, Daniel J Klein, Russell M Viner, and Chris Bonell. 2020. Determining the optimal strategy for reopening schools, the impact of test and trace interventions, and the risk of occurrence of a second COVID-19 epidemic wave in the UK: a modelling study. *The Lancet Child & Adolescent Health* 4, 11 (2020), 817–827.

[123] J. Panovska-Griffiths, C. C. Kerr, W. Waites, R. M. Stuart, D. Mistry, D. Foster, D. J. Klein, R. M. Viner, and C. Bonell. 2021. Modelling the potential impact of mask use in schools and society on COVID-19 control in the UK. *Scientific Reports* 11, 1 (2021), 8747.

[124] J. Panovska-Griffiths, R. M. Stuart, C. C. Kerr, K. Rosenfield, D. Mistry, W. Waites, D. J. Klein, C. Bonell, and R. M. Viner. 2022. Modelling the impact of reopening schools in the UK in early 2021 in the presence of the alpha variant and with roll-out of vaccination against SARS-CoV-2. *J. Math. Anal. Appl.* (2022), 126050.

[125] J. Panovska-Griffiths, B. Swallow, R. Hinch, J. Cohen, K. Rosenfeld, R. M. Stuart, L. Ferretti, F. Di Lauro, C. Wymant, A. Izzo, W. Waites, R. Viner, C. Bonell,

The COVID-19 Genomics UK (COG-UK) consortium, C. Fraser, D. Klein, and C. C. Kerr. 2022. Statistical and agent-based modelling of the transmissibility of different SARS-CoV-2 variants in England and impact of different interventions. *medRxiv* (2022). https://www.medrxiv.org/content/early/2022/01/03/2021.12.30.21267090

[126] Martin Pavelka, Kevin Van-Zandvoort, Sam Abbott, Katharine Sherratt, Marek Majdan, CMMID COVID-19 working group, Inštitút Zdravotnỳch Analỳz, Pavol Jarčuška, Marek Krajčí, Stefan Flasche, and Sebastian Funk. 2021. The impact of population-wide rapid antigen testing on SARS-CoV-2 prevalence in Slovakia. *Science* 372, 6542 (2021), 635–641.

[127] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[128] Patrick Peretti-Watel, Valérie Seror, Sébastien Cortaredona, Odile Launay, Jocelyn Raude, Pierrea Verger, Lisa Fressard, François Beck, Stéphane Legleye, Olivier l'Haridon, et al. 2020. A future vaccination campaign against COVID-19 at risk of vaccine hesitancy and politicisation. *The Lancet Infectious Diseases* (2020), 769–770.

[129] Liliana Perez and Suzana Dragicevic. 2009. An agent-based approach for modeling dynamics of contagious disease spread. *International journal of health geographics* 8, 1 (2009), 1–17.

[130] Hilary Piedrahita-Valdés, Diego Piedrahita-Castillo, Javier Bermejo-Higuera, Patricia Guillem-Saiz, Juan Ramón Bermejo-Higuera, Javier Guillem-Saiz, Juan Antonio Sicilia-Montalvo, and Francisco Machío-Regidor. 2021. Vaccine hesitancy on social media: Sentiment analysis from June 2011 to April 2019. *Vaccines* 9, 1 (2021), 28.

[131] Koen B Pouwels, Emma Pritchard, Philippa C Matthews, Nicole Stoesser, David W Eyre, Karina-Doris Vihta, Thomas House, Jodie Hay, John I Bell, John N Newton, et al. 2021. Effect of Delta variant on viral burden and vaccine effectiveness against new SARS-CoV-2 infections in the UK. *Nature Medicine* 27, 12 (2021), 2127–2135.

[132] Nataša Pržulj. 2007. Biological network comparison using graphlet degree distribution. *Bioinformatics* 23, 2 (2007), e177–e183.

[133] Natasa Pržulj, Derek G Corneil, and Igor Jurisica. 2004. Modeling interactome: scale-free or geometric? *Bioinformatics* 20, 18 (2004), 3508–3515.

[134] Jacek Radzikowski, Anthony Stefanidis, Kathryn H Jacobsen, Arie Croitoru, Andrew Crooks, and Paul L Delamater. 2016. The measles vaccination narrative in Twitter: a quantitative analysis. *JMIR Public Health and Surveillance* 2, 1 (2016), e5059.

[135] William M Rand. 1971. Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* 66, 336 (1971), 846–850.

[136] Erzsébet Ravasz and Albert-László Barabási. 2003. Hierarchical organization in complex networks. *Physical Review E* 67, 2 (2003), 026112.

[137] Mohammad S Razai, Pippa Oakeshott, Aneez Esmail, Charles Shey Wiysonge, Kasisomayajula Viswanath, and Melinda C Mills. 2021. COVID-19 vaccine hesitancy: the five Cs to tackle behavioural and sociodemographic factors. *Journal of the Royal Society of Medicine* 114, 6 (2021), 295–298. https://doi.org/10.1177/01410768211018951 arXiv:https://doi.org/10.1177/01410768211018951 PMID: 34077688.

[138] Pedro Ribeiro, David Aparício, Pedro Paredes, and Fernando Silva. 2017. GTScanner - Quick Discovery of Network Motifs. http://www.dcc.fc.up.pt/~daparicio/software. (2017). Accessed: 2019-08-25.

[139] Pedro Ribeiro, Pedro Paredes, Miguel E. P. Silva, David Aparicio, and Fernando Silva. 2021. A survey on subgraph counting: concepts, algorithms, and applications to network motifs and graphlets. *Comput. Surveys* 54, 2, Article 28 (March 2021), 36 pages. https://doi.org/10.1145/3433652

[140] Pedro Ribeiro and Fernando Silva. 2014. G-tries: a data structure for storing and finding subgraphs. *Data Mining and Knowledge Discovery* 28, 2 (2014), 337–377.

[141] David A Rier. 2007. The impact of moral suasion on internet HIV/AIDS support groups: evidence from a discussion of seropositivity disclosure ethics. *Health Sociology Review* 16, 3-4 (2007), 237–247.

[142] Matei Ripeanu, Ian Foster, and Adriana Iamnitchi. 2002. Mapping the gnutella network: properties of large-scale peer-to-peer systems and implications for system design. *arXiv preprint cs/0209028* (2002).

[143] Martin Ritchie, Luc Berthouze, Thomas House, and Istvan Z. Kiss. 2014. Higher-order structure and epidemic dynamics in clustered networks. *Journal of Theoretical Biology* 348 (2014), 21–32.

[144] Martin Ritchie, Luc Berthouze, and Istvan Z. Kiss. 2016. Generation and analysis of networks with a prescribed degree sequence and subgraph family: higher-order structure matters. *Journal of Complex Networks* 5, 1 (2016), 1–31.

[145] Justine Roberts. 2022. Mumsnet. `www.mumsnet.com`. (2022). Accessed: 01-05-2022.

[146] Elaine Robertson, Kelly S. Reeve, Claire L. Niedzwiedz, Jamie Moore, Margaret Blake, Michael Green, Srinivasa Vittal Katikireddi, and Michaela J. Benzeval. 2021. Predictors of COVID-19 vaccine hesitancy in the UK household longitudinal study. *Brain, Behavior, and Immunity* 94 (2021), 41–50.

[147] Alejandro Rodríguez-González, Juan Manuel Tuñas, Lucia Prieto Santamaría, Diego Fernández Peces-Barba, Ernestina Menasalvas Ruiz, Almudena Jaramillo, Manuel Cotarelo, Antonio J Conejo Fernández, Amalia Arce, and Angel Gil. 2020. Identifying polarity in tweets from an imbalanced dataset about diseases and vaccines using a meta-model based on machine learning techniques. *Applied Sciences* 10, 24 (2020), 9019.

[148] Lotte Romijn, Breanndán Ó Nualláin, and Leen Torenvliet. 2015. Discovering motifs in real-world social networks. In *International Conference on Current Trends in Theory and Practice of Informatics*. Springer, 463–474.

[149] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 1998. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision*. IEEE, 59–66.

[150] Marcel Salathé and Shashank Khandelwal. 2011. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS Computational Biology* 7, 10 (2011), e1002199.

[151] Marcel Salathé, Duy Q Vu, Shashank Khandelwal, and David R Hunter. 2013. The dynamics of health behavior sentiments on a large online social network. *EPJ Data Science* 2, 1 (2013), 1–12.

[152] Anida Sarajlić, Noël Malod-Dognin, Ömer Nebil Yaveroğlu, and Nataša Pržulj. 2016. Graphlet-based characterization of directed networks. *Scientific Reports* 6, 1 (2016), 1–14.

[153] Denise Scannell, Linda Desens, Marie Guadagno, Yolande Tra, Emily Acker, Kate Sheridan, Margo Rosner, Jennifer Mathieu, and Mike Fulk. 2021. COVID-19 vaccine discourse on Twitter: A content analysis of persuasion techniques, sentiment and mis/disinformation. *Journal of Health Communication* 26, 7 (2021), 443–459.

[154] Wolfgang E Schlauch and Katharina A Zweig. 2015. Influence of the null-model on motif detection. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 514–519.

[155] Ana Lucía Schmidt, Fabiana Zollo, Antonio Scala, Cornelia Betsch, and Walter Quattrociocchi. 2018. Polarization of the vaccination debate on Facebook. *Vaccine* 36, 25 (2018), 3606–3612. https://www.sciencedirect.com/science/article/pii/S0264410X18306601

[156] Benno Schwikowski, Peter Uetz, and Stanley Fields. 2000. A network of protein–protein interactions in yeast. *Nature Biotechnology* 18, 12 (2000), 1257–1261.

[157] Erin R Shellman, Charles F Burant, and Santiago Schnell. 2013. Network motifs provide signatures that characterize metabolism. *Molecular BioSystems* 9, 3 (2013), 352–360.

[158] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. 2002. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics* 31, 1 (2002), 64.

[159] Jonathon Shlens. 2014. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100* (2014).

[160] Constantinos Siettos, Cleo Anastassopoulou, Lucia Russo, Christos Grigoras, and Eleftherios Mylonakis. 2015. Modeling the 2014 ebola virus epidemic–agent-based simulations, temporal analysis and future predictions for liberia and sierra leone. *PLoS currents* 7 (2015).

[161] Miguel EP Silva, Pedro Paredes, and Pedro Ribeiro. 2017. Network motifs detection using random networks with prescribed subgraph frequencies. In *International Workshop on Complex Networks*. Springer, 17–29.

[162] Petrônio CL Silva, Paulo VC Batista, Hélder S Lima, Marcos A Alves, Frederico G Guimarães, and Rodrigo CP Silva. 2020. COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos, Solitons & Fractals* 139 (2020), 110088.

[163] Anika Singanayagam, Seran Hakki, and Jake Dunning et al. 2022. Community transmission and viral load kinetics of the SARS-CoV-2 delta (B.1.617.2) variant in vaccinated and unvaccinated individuals in the UK: a prospective, longitudinal, cohort study. *The Lancet Infectious Diseases* 22, 2 (2022), 183–195. https://doi.org/10.1016/S1473-3099(21)00648-4

[164] Rohit Singh, Jinbo Xu, and Bonnie Berger. 2008. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences* 105, 35 (2008), 12763–12768.

[165] Zoë C. Skea, Vikki A. Entwistle, Ian Watt, and Elizabeth Russell. 2008. 'Avoiding harm to others' considerations in relation to parental measles, mumps and rubella (MMR) vaccination discussions – An analysis of an online chat forum. *Social Science & Medicine* 67, 9 (2008), 1382–1390. https://doi.org/10.1016/j.socscimed.2008.07.006

[166] Maria Skeppstedt, Andreas Kerren, and Manfred Stede. 2017. Automatic detection of stance towards vaccination in online discussion forums. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*. 1–8.

[167] Maria Skeppstedt, Andreas Kerren, and Manfred Stede. 2018. Vaccine hesitancy in discussion forums: computer-assisted argument mining with topic models. In *MIE*. 366–370.

[168] Olaf Sporns, Rolf Kötter, and Karl J Friston. 2004. Motifs in brain networks. *PLoS Biology* 2, 11 (2004), e369.

[169] Mattia Tantardini, Francesca Ieva, Lucia Tajoli, and Carlo Piccardi. 2019. Comparing methods for comparing networks. *Scientific Reports* 9, 1 (2019), 1–19.

[170] Lara Tavoschi, Filippo Quattrone, Eleonora D'Andrea, Pietro Ducange, Marco Vabanesi, Francesco Marcelloni, and Pier Luigi Lopalco. 2020. Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from September 2016 to August 2017 in Italy. *Human vaccines & immunotherapeutics* 16, 5 (2020), 1062–1069.

[171] Robin N Thompson, T Déirdre Hollingsworth, Valerie Isham, Daniel Arribas-Bel, Ben Ashby, Tom Britton, Peter Challenor, Lauren HK Chappell, Hannah Clapham, Nik J Cunniffe, et al. 2020. Key questions for modelling COVID-19 exit strategies. *Proceedings of the Royal Society B* 287, 1932 (2020), 20201405.

[172] Melissa Tracy, Magdalena Cerdá, and Katherine M Keyes. 2018. Agent-based modeling in public health: current applications and future directions. *Annual review of public health* 39 (2018), 77.

[173] John S Tregoning, Katie E Flight, Sophie L Higham, Ziyin Wang, and Benjamin F Pierce. 2021. Progress of the COVID-19 vaccine effort: viruses, vaccines and variants versus efficacy, effectiveness and escape. *Nature Reviews Immunology* 21, 10 (2021), 626–636.

[174] United Nations. 2019. Household Size and Composition. (2019). `https://www.un.org/development/desa/pd/data/household-size-and-composition`

[175] Sergi Valverde and Ricard V Solé. 2005. Network motifs in computational graphs: A case study in software architecture. *Physical Review E* 72, 2 (2005), 026107.

[176] Alexei Vázquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. 2003. Modeling of protein interaction networks. *Complexus* 1, 1 (2003), 38–44.

[177] Russell Viner, Simon Russell, and Rosella Saulle. 2022. School Closures During Social Lockdown and Mental Health, Health Behaviors, and Well-being Among

Children and Adolescents During the First COVID-19 Wave: A Systematic Review. *JAMA Pediatrics* 176 (2022), 400–409. Issue 4.

[178] Mingda Wang and Guangmin Hu. 2020. A novel method for twitter sentiment analysis based on attentional-graph neural network. *Information* 11, 2 (2020), 92.

[179] Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of 'small-world'networks. *Nature* 393, 6684 (1998), 440–442.

[180] Anatol E Wegner, Luis Ospina-Forero, Robert E Gaunt, Charlotte M Deane, and Gesine Reinert. 2018. Identifying networks with common organizational principles. *Journal of Complex Networks* 6, 6 (2018), 887–913.

[181] R Weller, S White, et al. 2016. A content analysis of online forum discussion about measles, mumps and rubella (MMR) vaccination between 2004 and 2015. *International Journal of Pharmacy Practice* 24 (2016), 25–25.

[182] Gerhard Widmer and Miroslav Kubat. 1996. Learning in the presence of concept drift and hidden contexts. *Machine Learning* 23, 1 (1996), 69–101.

[183] Lander Willem, Frederik Verelst, Joke Bilcke, Niel Hens, and Philippe Beutels. 2017. Lessons from a decade of individual-based models for infectious disease transmission: a systematic review (2006-2015). *BMC infectious diseases* 17, 1 (2017), 1–16.

[184] Jacqui Wise. 2021. Covid-19: rare immune response may cause clots after AstraZeneca vaccine, say researchers. *BMJ* 373 (2021). `https://doi.org/10.1136/bmj.n954` arXiv:https://www.bmj.com/content/373/bmj.n954.full.pdf

[185] Achim Wolf, Jack Hulmes, and Susan Hopkins. 2021. Lateral flow device specificity in phase 4 (post marketing) surveillance. (2021).

[186] Christopher Wolfram. 2020. An agent-based model of covid-19. *Complex Syst* 29, 1 (2020), 87–105.

[187] World Health Organisation (WHO). 2019. Ten threats to global health in 2019. (2019). https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019.

[188] World Health Organization. 2022. Antigen-detection in the diagnosis of SARS-CoV-2 infection: Interim guidance 6 October 2021. (2022). https://www.who.int/publications/i/item/antigen-detection-in-the-diagnosis-of-sars-cov-2infection-using-rapid-immunoassays WHO Reference Number: WHO/2019-nCoV/Antigen_Detection/2021.1.

[189] World Health Organization. 2022. Tracking SARS-CoV-2 variants. (2022). https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/

[190] Xiaochuan Xu and Gesine Reinert. 2018. Triad-based comparison and signatures of directed networks. *International Conference on Complex Networks and their Applications* (2018), 590–602.

[191] Jaewon Yang and Jure Leskovec. 2015. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems* 42, 1 (2015), 181–213.

[192] Ömer Nebil Yaveroğlu, Noël Malod-Dognin, Darren Davis, Zoran Levnajic, Vuk Janjic, Rasa Karapandza, Aleksandar Stojmirovic, and Nataša Pržulj. 2014. Revealing the hidden language of complex networks. *Scientific Reports* 4, 1 (2014), 1–9.

[193] Ömer Nebil Yaveroğlu, Tijana Milenković, and Nataša Pržulj. 2015. Proper evaluation of alignment-free network comparison methods. *Bioinformatics* 31, 16 (2015), 2697–2704.

[194] Esti Yeger-Lotem, Shmuel Sattath, Nadav Kashtan, Shalev Itzkovitz, Ron Milo, Ron Y Pinter, Uri Alon, and Hanah Margalit. 2004. Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction. *Proceedings of the National Academy of Sciences* 101, 16 (2004), 5934–5939.

[195] Bernadette C Young, David W Eyre, Saroj Kendrick, Chris White, Sylvester Smith, George Beveridge, Toby Nonnenmacher, Fegor Ichofu, Joseph Hillier, Sarah Oakley, Ian Diamond, Emma Rourke, Fiona Dawe, Ieuan Day, Lisa Davies, Paul Staite, Andrea Lacey, James McCrae, Ffion Jones, Joseph Kelly, Urszula Bankiewicz, Sarah Tunkel, Richard Ovens, David Chapman, Vineta Bhalla, Peter Marks, Nick Hicks, Tom Fowler, Susan Hopkins, Lucy Yardley, and Tim E A Peto. 2021. Daily testing for contacts of individuals with SARS-CoV-2 infection

and attendance and SARS-CoV-2 transmission in English secondary schools and colleges: an open-label, cluster-randomised trial. *The Lancet* 398, 10307 (2021), 1217–1229.

[196] Xiaoyi Yuan and Andrew T. Crooks. 2018. Examining online vaccination discussion and communities in twitter. In *Proceedings of the 9th International Conference on Social Media and Society*. 197–206. https://doi.org/10.1145/3217804.3217912

[197] Xiaoyi Yuan, Ross J Schuchard, and Andrew T Crooks. 2019. Examining emergent communities and social bots within the polarized online vaccination debate in Twitter. *Social media+ society* 5, 3 (2019), 2056305119865465.

[198] Wayne W Zachary. 1977. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* 33, 4 (1977), 452–473.

[199] Kou Zhongbao and Zhang Changshui. 2003. Reply networks on a bulletin board system. *Physical Review E* 67, 3 (2003), 036117.

[200] Dongxiao Zhu and Zhaohui S Qin. 2005. Structural comparison of metabolic networks in selected single cell organisms. *BMC bioinformatics* 6, 1 (2005), 1–12.

Appendix A

NetEmd with different metrics

A.1 Area Under Precision-Recall (AUPR)

Sarajlić et al. [152] evaluate the performance of their graph comparison tool using the area under precision-recall curve (AUPR), after the framework proposed by Yaveroğlu et al. [193]. The metric is calculated as follows: for each value of a parameter $\varepsilon \geq 0$, if the distance between two networks is less than ε, then the networks belong to the same cluster. The parameter ε ranges from 0 to 1, with values incrementing by 5×10^{-3}. For each value of ε, we compute the number of *true positives*, *false positives*, *true negatives* and *false negatives*. In this context, we define these concepts as:

- *True Positive (TP)*: the distance between the networks is smaller than ε and they were generated by the same random network model.
- *False Positive (FP)*: the distance between the networks is smaller than ε but they were generated by different random network models.
- *True Negative (TN)*: the distance between the networks is greater than ε and they were generated by different random network models.
- *False Negative (FN)*: the distance between the networks is greater than ε but they were generated by the same random network model.

From these quantities, we calculate *precision* as $\frac{TP}{TP+FP}$ and *recall* as $\frac{TP}{TP+FN}$, obtaining 200 tuples of (precision, recall), from which we are able to plot the precision-recall curve. The area under this curve is a metric relevant to our problem as it puts an emphasis on the positive predictive value of the model, disregarding the true negatives

that compose the majority of datasets with imbalanced labels such as this one (with 8 random network models, positive labels comprise only 12.5% of the dataset).

A.1.1 Undirected Results

Table B.5 shows the area under precision-recall curve (AUPR) for Task 1 and Task 2 in the undirected case and the Onnela et al. dataset.

In Task 1, according to AUPR, using Principal Component Analysis (PCA) or Independent Component Analysis (ICA) does not lead to a gain in performance, unlike what we observed when using $\overline{P}$. This is not surprising as the improvement we displayed with $\overline{P}$ was within a standard error and both methodologies are able to separate networks between these random models with a very high degree of accuracy. Task 1 serves as a proof of concept that our proposed measure is able to perform the most simple task.

The results for Task 2 are aligned with what we observe when using $\overline{P}$, coupling ICA with NetEmd leads to the best separation between models when different network sizes and densities are present. In the Onnela et al. dataset, the results with AUPR also agree with $\overline{P}$, both performance metrics consider the original NetEmd with size 4 orbits to be one of the top performers with this dataset. The difference between the two metrics is that when measuring with $\overline{P}$, we observe that using PCA with 90% explained variance leads to similar performance (Table B.2), but the top performer in this dataset according to AUPR is PCA with 99% explained variance, achieving an AUPR score of 0.791.

A.1.2 Directed Results

The results using AUPR for Task 1 in directed networks are shown in Table B.6 and for Task 2 and the dataset of real world directed networks in Table B.7.

In Task 1, the main difference we find is that Graphlet Degree Distribution Agreement (GDA) with size 3 orbits shows an improvement over the other algorithms, becoming the top performer with 25% (tied with Directed Graphlet Correlation Degree (DGCD) with 129 orbits) and 75% reciprocity. DGCD with 129 orbits still shows the best performance with 0% reciprocity, but PCA with 90, 95 and 99% explained variance is within one standard error, similarly to the results measured with $\overline{P}$. At a 100% reciprocity, NetEmd with size 4 orbits achieves the best performance, but PCA with 90, 95 and 99% explained variance and ICA with 2 components obtain similar AUPR

scores. As before, we highlight that the high scores achieved by the 6 algorithms in Table B.6 serve as a proof of concept that they are able to distinguish the random network models for different levels of reciprocity, even if the AUPR score is not within a standard error of the best score.

In Task 2, similarly to the undirected version, we find an agreement between the AUPR score and $\overline{P}$, with ICA with 2 components highlighted as the highest performing algorithm for 0, 25, 50 and 75% reciprocity datasets and NetEmd with size 4 orbits for 100% reciprocity. The same agreement between AUPR and $\overline{P}$ is found on the dataset of real world directed networks, with the top performer being ICA with 6 components with an AUPR of 0.891 but not highlighted in Table B.7.

The conclusions regarding the difference between our proposed NetEmd methods and TriadEMD also hold when using AUPR instead of $\overline{P}$, in the synthetic datasets TriadEMD achieves better AUPR scores that both NetEmd and *Weighted_NetEmd* with orbits from size up to 3 graphlets, but worse results than NetEmd with size 4 graphlets. In the real world directed networks dataset, both NetEmd and *Weighted_NetEmd* with size 3 orbits obtain better AUPR score than TriadEMD.

A.2 Adjusted Rand Index

The adjusted Rand index (ARI) [70, 135] is a metric to evaluate the similarity of data clusterings; a corrected for chance version of the Rand index [135] was proposed by Hubert and Arabie [70]. Let $\mathcal{G} = \{G_1, G_2, \ldots, G_N\}$ be the set of networks we wish to partition into clusters. For the synthetic datasets, the ground truth partition of this set is known, we have 8 models of random graphs and we evaluate a network comparison method by its ability to reconstruct these groups. For the Onnela et al. [118] dataset, the ground truth partition used is the one shown in the supplementary material of [4]. Let $\mathcal{P}^* = \{P_1^*, P_2^*, \ldots, P_c^*\}$ be the ground truth partitions and $\mathcal{P} = \{P_1, P_2, \ldots, P_c\}$ be the clusters produced by the network comparison algorithms. These clusters are generated by applying a hierarchical clustering algorithm to the matrix of pair-wise distances calculated by the network comparison; in our case we used the implementation by *scikit-learn* [127], the *AgglomerativeClustering* class with complete linkage.

Given the two set of clusters, we define the following four quantities:

- a: the number of pairs of elements of $\mathcal{G}$ that are in the same cluster in $\mathcal{P}^*$ and in the same cluster in $\mathcal{P}$.

- b: the number of pairs of elements of $\mathcal{G}$ that are in different clusters in $\mathcal{P}^*$ and also in different clusters in $\mathcal{P}$.
- c: the number of pairs of elements of $\mathcal{G}$ that are in the same cluster in $\mathcal{P}^*$ but in different clusters in $\mathcal{P}$.
- d: the number of pairs of elements of $\mathcal{G}$ that are in different clusters in $\mathcal{P}^*$ but in the same cluster in $\mathcal{P}$.

These quantities can be interpreted as the number of true positives, true negatives, false negatives and false positives, respectively, if we consider the attribution of a pair of networks to a cluster as a decision problem. The Rand index is then defined as:

$$RI = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{N}{2}}.$$

The general form for correcting a index for chance is:

$$\frac{RI - \mathbb{E}(RI)}{\max(RI) - \mathbb{E}(RI)}.$$

Hubert and Arabie [70] show how to calculate $\mathbb{E}(RI)$, which we omit for brevity, leading to the formula for the adjusted Rand index (assuming a maximum value of 1 for the Rand index when c and d are 0):

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}{\binom{N}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}{\binom{N}{2}}},$$

where n_{ij} is the number of networks in common between P_i^* and P_j, i.e., $n_{ij} = |P_i^* \cap P_j|$, $a_i = \sum_{j=1}^{c} n_{ij}$ and $b_j = \sum_{i=1}^{c} n_{ij}$.

A.2.1 Undirected Results

Table B.8 shows the adjusted Rand index (ARI) for Task 1, Task 2 and real world networks in the undirected case.

As with AUPR, we find that using ARI as a performance metric does not change the conclusions drawn from the results with $\overline{P}$. Recalling that the ARI metric measures the quality of hierarchical clusters generated from a distance matrix, for Task 1, the output of original NetEmd with size 5 orbits leads the best score using this metric.

More evidence of the quality of the clusters produced by the original NetEmd with size 5 orbits is to notice that increasing the number of components in ICA or the percentage of variance explained in PCA (and therefore approximating the results of original NetEmd) leads to higher ARI score.

For Task 2, the gain measured by $\overline{P}$ and AUPR when using *ICA_NetEmd* with 2 components is also reflected in the ARI score. For the Onnela et al. dataset, the best performance we measure not included in Table B.8 is *ICA_NetEmd* with size 4 graphlets and 8 components, with the components calculated with size 5 graphlets, with a score of 0.713.

A.2.2 Directed Results

The results using ARI for Task 1 in directed networks are shown in Table B.9 and for Task 2 and the dataset of real world directed networks in Table B.10.

In Task 1, we find that, according to ARI, the performance of all NetEmd variants is significantly below the performance of DGCD with 129 orbits for 0, 25 and 50% reciprocity and of GDA with size 3 graphlets for 75% reciprocity. At 100% reciprocity, NetEmd, *PCA_NetEmd* with 80% variance and GDA, all with size 4 graphlets, share the top performance, but with a smaller difference to other algorithms and parameters. These results are a stark contrast to the ones we observe using AUPR and $\overline{P}$, where scores for this task did not differ significantly between all the algorithms.

In Task 2, unlike in the other metrics, *ICA_NetEmd* with 2 components is no longer considered the best performance for 0%, with TriadEMD achieving the best result instead. On the other hand, according to ARI, *ICA_NetEmd* with 2 components is the top performer for 100% reciprocity. For 25, 50 and 75% reciprocity, *ICA_NetEmd* with 2 components performs significantly worse when comparing to the results using AUPR and $\overline{P}$. ARI scores indicate that the best performers in these levels of reciprocity are *PCA_NetEmd* with 99% variance, NetEmd with size 3 graphlets and *PCA_NetEmd* with 95% variance. As we mention previously, using a smaller graphlet size usually leads to better performance when different network sizes and densities are involved. We find partial agreement to this claim when using ARI as the performance metric. At 0% reciprocity, we find that *ICA_NetEmd* with 2 components and size 3 orbits achieves an ARI of 0.509 (higher than the score of *ICA_NetEmd* with 2 components and size 4 orbits in Table B.10, but lower than the score of TriadEMD); at 75% reciprocity, *PCA_NetEmd* with 90% variance explained and size 3 orbits achieves an ARI of 0.495 (higher than the score of *PCA_NetEmd* with 95% variance explained and size

4 orbits in Table B.10).

For the dataset of real world directed networks, ARI disagrees with AUPR and $\overline{P}$, assigning the best performance to GDA with size 3 orbits. This task also favours using smaller graphlets for the comparison, so the *PCA_NetEmd* and *ICA_NetEmd* results are at a disadvantage compared to the other methods as we only report results with size 4 in Table B.10. For these two algorithms with size 3 graphlets, the best ARI scores on this task are 0.524, when using *PCA_NetEmd* with 90% explained variance, and 0.489, when using *ICA_NetEmd* with 19 components.

On the overall, in spite of some agreements between ARI and AUPR or $\overline{P}$, we find ARI to be an unreliable metric to gauge the performance of network distance measures, as it relies on an intermediate step of clustering that is unrelated to the measures themselves. Using a different clustering method or even a different linkage leads to disparate results compared to the ones we present in this section, which adds a confounding factor when measuring the performance of the network comparison itself. An example of this unreliability is the ARI score of 0 given to GDA with size 3 orbits and to TriadEMD at 100% reciprocity, when the other metrics showed a performance comparable to DGCD.

Appendix B

NetEmd Results tables

Algorithm	Parameter	Reciprocity				
		0%	25%	50%	75%	100%
All Orbits	G3D	0.989 ± 0.002	0.990 ± 0.002	0.989 ± 0.002	0.991 ± 0.002	0.968 ± 0.003
	G4D	0.992 ± 0.002	0.993 ± 0.001	0.993 ± 0.001	*0.993 ± 0.001*	*0.989 ± 0.001*
Weighted_NetEmd	G3D	0.989 ± 0.002	0.990 ± 0.002	0.989 ± 0.002	0.989 ± 0.002	0.969 ± 0.003
	G4D	0.992 ± 0.002	0.993 ± 0.001	0.992 ± 0.001	*0.993 ± 0.001*	*0.989 ± 0.001*
PCA_NetEmd	50% Variance	0.989 ± 0.003	0.993 ± 0.002	0.991 ± 0.002	*0.993 ± 0.001*	0.978 ± 0.004
	80% Variance	0.990 ± 0.003	0.993 ± 0.001	0.990 ± 0.002	0.992 ± 0.001	0.982 ± 0.003
	90% Variance	*0.993 ± 0.002*	0.993 ± 0.001	0.992 ± 0.001	*0.993 ± 0.001*	**0.990 ± 0.002**
	95% Variance	0.992 ± 0.002	0.993 ± 0.001	0.992 ± 0.001	*0.993 ± 0.001*	*0.988 ± 0.001*
	99% Variance	*0.992 ± 0.002*	0.992 ± 0.001	0.992 ± 0.001	0.992 ± 0.001	0.986 ± 0.002
ICA_NetEmd	2 components	0.990 ± 0.003	0.993 ± 0.002	0.992 ± 0.002	*0.993 ± 0.001*	0.987 ± 0.002
	10 components	*0.992 ± 0.002*	0.993 ± 0.001	0.992 ± 0.002	*0.993 ± 0.001*	0.985 ± 0.002
	15 components	*0.993 ± 0.002*	0.994 ± 0.001	0.993 ± 0.001	*0.993 ± 0.001*	0.980 ± 0.001
TriadEMD		0.992 ± 0.001	0.991 ± 0.001	0.990 ± 0.002	0.992 ± 0.002	0.970 ± 0.003
DGCD	13 orbits	0.986 ± 0.003	0.992 ± 0.002	0.990 ± 0.002	0.988 ± 0.002	0.968 ± 0.004
	129 orbits	**0.995 ± 0.002**	**0.996 ± 0.001**	**0.995 ± 0.001**	**0.995 ± 0.002**	0.987 ± 0.002
GDA	G3D	0.988 ± 0.003	0.990 ± 0.002	0.988 ± 0.003	0.992 ± 0.002	0.959 ± 0.005
	G4D	0.992 ± 0.003	0.992 ± 0.002	0.991 ± 0.003	*0.993 ± 0.002*	0.977 ± 0.007

Table B.1: Results for Task 1 in directed networks, for each level of reciprocity. The metric used for evaluation is the mean (and standard error of the mean) of the 16 values for $\overline{P}$, in each combination of number of nodes, network density and reciprocity. The parameter column indicates: graphlet size used in directed and weighted NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 4); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 4; the number of orbits used by Directed Graphlet Correlation Degree (DGCD); graphlet sizes used in Graphlet Degree Distribution Agreement (GDA). The bolded values are the maximum for each reciprocity level. Italic values are within one standard error of the maximum performance.

Parameters	Synthetic		Real
	Task 1	Task 2	Onnela et al.
3 - 50% Variance	0.981 ± 0.004	0.912	0.872
3 - 80% Variance	0.980 ± 0.003	0.924	0.874
3 - 90% Variance	0.984 ± 0.002	0.929	0.881
3 - 95% Variance	0.982 ± 0.002	0.911	0.891
3 - 99% Variance	0.986 ± 0.002	0.928	0.889
4 - 50% Variance	0.986 ± 0.003	0.917	0.873
4 - 80% Variance	0.987 ± 0.003	0.926	0.885
4 - 90% Variance	0.992 ± 0.001	0.931	**0.898**
4 - 95% Variance	0.990 ± 0.001	0.921	0.897
4 - 99% Variance	0.991 ± 0.001	0.931	0.897
5 - 50% Variance	0.986 ± 0.003	0.915	0.863
5 - 80% Variance	0.985 ± 0.002	0.918	0.893
5 - 90% Variance	*0.995 ± 0.001*	0.922	0.891
5 - 95% Variance	0.993 ± 0.001	0.919	0.892
5 - 99% Variance	*0.995 ± 0.001*	0.921	0.891
Other best	**0.996 ± 0.001** (ICA G5 13 components)	**0.933** (ICA G5 7 components)	**0.898** (Original NetEmd G4)

Table B.2: Results for *PCA_NetEmd* with different combinations of graphlet size and percentage of variance explained in Task 1, Task 2 and the Onnela et al. dataset in the undirected case. The metric used for Task 1 is the mean (and standard error of the mean) of the 16 values for $\overline{P}$ in each combination of number of nodes with network density. The metric used for Task 2 is the sole value of $\overline{P}$ after comparing the 1280 networks. The bolded values are the maximum in each task. Italic values are within one standard error of the maximum performance.

Parameter	Reciprocity				
	0%	25%	50%	75%	100%
3 - 50% Variance	0.984 ± 0.003	0.987 ± 0.003	0.985 ± 0.003	0.987 ± 0.003	0.928 ± 0.004
3 - 80% Variance	0.985 ± 0.003	0.991 ± 0.002	0.991 ± 0.002	0.992 ± 0.002	0.932 ± 0.004
3 - 90% Variance	0.989 ± 0.003	0.991 ± 0.002	0.991 ± 0.002	0.990 ± 0.002	0.929 ± 0.005
3 - 95% Variance	0.992 ± 0.002	0.992 ± 0.001	0.990 ± 0.002	0.992 ± 0.002	0.940 ± 0.003
3 - 99% Variance	0.990 ± 0.002	0.991 ± 0.001	0.990 ± 0.002	0.991 ± 0.002	0.938 ± 0.004
4 - 50% Variance	0.989 ± 0.003	0.993 ± 0.002	0.991 ± 0.002	*0.993 ± 0.001*	0.978 ± 0.004
4 - 80% Variance	0.990 ± 0.003	0.993 ± 0.001	0.990 ± 0.002	0.992 ± 0.001	0.982 ± 0.003
4 - 90% Variance	*0.993 ± 0.002*	0.993 ± 0.001	0.992 ± 0.001	*0.993 ± 0.001*	**0.990 ± 0.002**
4 - 95% Variance	0.992 ± 0.002	0.993 ± 0.001	0.992 ± 0.001	*0.993 ± 0.001*	*0.988 ± 0.001*
4 - 99% Variance	0.992 ± 0.002	0.992 ± 0.001	0.992 ± 0.001	0.992 ± 0.001	0.986 ± 0.002
Other best	**0.995 ± 0.002** (DGCD129)	**0.996 ± 0.001** (DGCD129)	**0.995 ± 0.001** (DGCD129)	**0.995 ± 0.002** (DGCD129)	*0.989 ± 0.001* (Weighted G4D)

Table B.3: Results for *PCA_NetEmd* with different combinations of graphlet size and percentage of variance explained in Task 1 of directed networks, for each level of reciprocity. The metric used for evaluation is the mean (and standard error of the mean) of the 16 values for $\overline{P}$, in each combination of number of nodes, network density and reciprocity. The bolded values are the maximum for each reciprocity level. Italic values are within one standard error of the maximum performance.

Parameters	Synthetic					Real Networks
	Reciprocity					
	0%	25%	50%	75%	100%	
3 - 50% Variance	0.915	0.919	0.918	0.918	0.857	0.825
3 - 80% Variance	0.925	0.909	0.905	0.907	0.865	0.858
3 - 90% Variance	0.933	0.909	0.905	0.906	0.866	0.861
3 - 95% Variance	0.934	0.908	0.905	0.904	0.867	0.868
3 - 99% Variance	0.921	0.903	0.900	0.903	0.879	0.867
4 - 50% Variance	0.924	0.897	0.900	0.894	0.908	0.821
4 - 80% Variance	0.924	0.885	0.892	0.881	0.920	0.825
4 - 90% Variance	0.921	0.882	0.889	0.878	**0.927**	0.824
4 - 95% Variance	0.915	0.880	0.887	0.876	0.922	0.822
4 - 99% Variance	0.914	0.878	0.885	0.875	0.924	0.818
Other best	**0.941** (ICA G4D 2 comps)	**0.926** (ICA G3D 2 comps)	**0.925** (ICA G3D 2 comps)	**0.928** (ICA G3D 2 comps)	**0.927** (ICA G4D 4 comps)	**0.870** (ICA G3D 10 comps)

Table B.4: Results for *PCA_NetEmd* with different combinations of graphlet size and percentage of variance explained in Task 2 in directed networks, for each level of reciprocity, and for the real world networks dataset. The metric used for evaluation is the mean (and standard error of the mean) of the 16 values for $\overline{P}$, in each combination of number of nodes, network density and reciprocity. The bolded values are the maximum for each reciprocity level.

Algorithm	Parameter	Synthetic		Real
		Task 1	Task 2	Onnela et al.
Original NetEmd	G3	0.91 ± 0.01	0.670	0.762
	G4	*0.95 ± 0.01*	0.705	**0.790**
	G5	**0.96 ± 0.01**	0.712	0.761
PCA	50% Variance	0.93 ± 0.01	0.668	0.702
	80% Variance	0.92 ± 0.01	0.701	0.759
	90% Variance	*0.96 ± 0.01*	0.717	0.756
	95% Variance	*0.96 ± 0.01*	0.716	0.760
	99% Variance	**0.96 ± 0.01**	0.718	0.760
ICA	2 components	0.93 ± 0.01	**0.742**	0.717
	10 components	0.94 ± 0.01	0.728	0.750
	15 components	**0.96 ± 0.01**	0.729	0.754
GCD	11 orbits	0.89 ± 0.02	0.516	0.698
	73 orbits	0.94 ± 0.02	0.600	0.747
GDA	G3	0.92 ± 0.02	0.605	0.783
	G4	0.80 ± 0.04	0.535	0.692
	G5	0.74 ± 0.04	-	0.638

Table B.5: Results for Task 1, Task 2 and the Onnela et al. [118] datasets in undirected networks. The metric used for Task 1 is the mean (and standard error of the mean) of the 16 values for area under precision-recall curve (AUPR) in each combination of number of nodes with network density. The metric used for Task 2 and Onnela et al. dataset is the sole value of AUPR after comparing the 1280 and 151 networks, respectively. The parameter column indicates: graphlet size used in original NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 5); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 5; the number of orbits used by Graphlet Correlation Degree (GCD); graphlet sizes used in GDA. The bolded values are the maximum for each task. Italic values are within one standard error of the maximum performance. Note that results for GDA with graphlet size 5 took longer than a week to return results for Task 2, at which point we stopped the computation.

Algorithm	Parameter	Reciprocity				
		0%	25%	50%	75%	100%
All Orbits	G3D	0.89 ± 0.01	0.90 ± 0.01	0.89 ± 0.01	0.90 ± 0.01	0.914 ± 0.007
	G4D	*0.961 ± 0.007*	0.945 ± 0.005	**0.941 ± 0.005**	0.944 ± 0.005	**0.992 ± 0.002**
Weighted_NetEmd	G3D	0.87 ± 0.01	0.88 ± 0.01	0.87 ± 0.02	0.89 ± 0.02	0.875 ± 0.009
	G4D	0.93 ± 0.01	0.930 ± 0.007	0.925 ± 0.009	0.928 ± 0.007	0.923 ± 0.009
PCA_NetEmd	50% Variance	0.94 ± 0.01	0.929 ± 0.009	0.928 ± 0.008	0.927 ± 0.008	0.957 ± 0.005
	80% Variance	0.94 ± 0.01	0.935 ± 0.006	0.932 ± 0.008	0.932 ± 0.006	0.958 ± 0.006
	90% Variance	*0.958 ± 0.008*	0.938 ± 0.005	*0.936 ± 0.005*	0.936 ± 0.005	0.974 ± 0.004
	95% Variance	*0.958 ± 0.007*	0.940 ± 0.005	*0.937 ± 0.005*	0.939 ± 0.005	0.971 ± 0.003
	99% Variance	*0.958 ± 0.007*	0.941 ± 0.005	*0.936 ± 0.005*	0.940 ± 0.005	0.977 ± 0.002
ICA_NetEmd	2 components	0.941 ± 0.007	0.938 ± 0.007	0.934 ± 0.007	0.940 ± 0.007	0.982 ± 0.003
	10 components	0.936 ± 0.009	0.941 ± 0.005	*0.939 ± 0.007*	0.937 ± 0.006	0.936 ± 0.005
	15 components	0.946 ± 0.007	0.938 ± 0.005	*0.938 ± 0.006*	0.937 ± 0.005	0.92 ± 0.01
TriadEMD		0.938 ± 0.006	0.936 ± 0.008	0.93 ± 0.01	0.94 ± 0.01	0.936 ± 0.005
DGCD	13 orbits	0.92 ± 0.02	0.93 ± 0.02	0.91 ± 0.02	0.89 ± 0.02	0.87 ± 0.02
	129 orbits	**0.97 ± 0.01**	**0.96 ± 0.01**	*0.94 ± 0.02*	0.92 ± 0.02	0.90 ± 0.02
GDA	G3D	0.90 ± 0.03	**0.96 ± 0.01**	*0.94 ± 0.01*	**0.96 ± 0.01**	0.82 ± 0.04
	G4D	0.88 ± 0.03	0.84 ± 0.03	0.86 ± 0.03	0.84 ± 0.03	0.80 ± 0.04

Table B.6: Results for Task 1 in directed networks, for each level of reciprocity. The metric used for evaluation is the mean (and standard error of the mean) of the 16 values for AUPR, in each combination of number of nodes, network density and reciprocity. The parameter column indicates: graphlet size used in directed and weighted NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 4); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 4; the number of orbits used by DGCD; graphlet sizes used in GDA. The bolded values are the maximum for each reciprocity level. Italic values are within one standard error of the maximum performance.

Algorithm	Parameter	Synthetic					Real Networks
		Reciprocity					
		0%	25%	50%	75%	100%	
All Orbits	G3D	0.644	0.594	0.583	0.598	0.721	**0.883**
	G4D	0.702	0.630	0.632	0.628	**0.953**	0.820
Weighted	G3D	0.605	0.554	0.545	0.559	0.639	**0.883**
	G4D	0.608	0.567	0.563	0.565	0.660	0.847
PCA	50% Variance	0.696	0.638	0.644	0.631	0.855	0.829
	80% Variance	0.700	0.626	0.634	0.621	0.845	0.824
	90% Variance	0.718	0.628	0.639	0.623	0.876	0.822
	95% Variance	0.705	0.628	0.636	0.623	0.872	0.822
	99% Variance	0.703	0.628	0.637	0.623	0.892	0.819
ICA	2 components	**0.767**	**0.688**	**0.691**	**0.683**	0.926	0.726
	10 components	0.725	0.669	0.679	0.664	0.820	0.837
	15 components	0.729	0.660	0.680	0.658	0.768	0.841
TriadEMD		0.664	0.629	0.623	0.629	0.738	0.881
DGCD	13 orbits	0.631	0.625	0.600	0.581	0.567	0.872
	129 orbits	0.653	0.643	0.620	0.593	0.558	0.860
GDA	G3D	0.584	0.552	0.539	0.551	0.549	0.760
	G4D	-	-	-	-	0.535	-

Table B.7: Results for Task 2 in directed networks, for each level of reciprocity, and for the real world networks dataset. The metric used for this task is the sole value of AUPR after comparing the 1280 and 1231 networks, respectively. The parameter column indicates: graphlet size used in directed and weighted NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 4); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 4; the number of orbits used by DGCD; graphlet sizes used in GDA. The bolded values are the maximum for each dataset. Note that results for GDA with graphlet size 4 took longer than a week to return results for Task 2 and in the real world networks dataset, at which point we stopped the computation.

Algorithm	Parameter	Synthetic		Real
		Task 1	Task 2	Onnela et al.
Original NetEmd	G3	0.63 ± 0.03	0.472	0.592
	G4	0.69 ± 0.02	0.427	**0.675**
	G5	$\mathbf{0.81 \pm 0.02}$	0.445	0.576
PCA	50% Variance	0.72 ± 0.03	0.439	0.558
	80% Variance	0.71 ± 0.03	0.447	0.619
	90% Variance	0.78 ± 0.02	0.436	0.612
	95% Variance	$\mathit{0.80 \pm 0.02}$	0.414	0.610
	99% Variance	$\mathbf{0.81 \pm 0.02}$	0.396	0.587
ICA	2 components	0.67 ± 0.02	**0.477**	0.521
	10 components	0.73 ± 0.03	0.452	0.593
	15 components	0.76 ± 0.03	0.469	0.602
GCD	11 orbits	0.64 ± 0.07	0.309	0.435
	73 orbits	0.77 ± 0.05	0.430	0.488
GDA	G3	0.74 ± 0.05	0.361	0.456
	G4	0.62 ± 0.07	0.332	0.197
	G5	0.41 ± 0.05	-	0.207

Table B.8: Results for Task 1, Task 2 and the Onnela et al. [118] datasets in undirected networks. The metric used for Task 1 is the mean (and standard error of the mean) of the 16 values for adjusted Rand index (ARI) in each combination of number of nodes with network density. The metric used for Task 2 and Onnela et al. dataset is the sole value of ARI after comparing the 1280 and 151 networks, respectively. The parameter column indicates: graphlet size used in original NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 5); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 5; the number of orbits used by GCD; graphlet sizes used in GDA. The bolded values are the maximum for each task. Italic values are within one standard error of the maximum performance. Note that results for GDA with graphlet size 5 took longer than a week to return results for Task 2, at which point we stopped the computation.

Algorithm	Parameter	Reciprocity				
		0%	25%	50%	75%	100%
All Orbits	G3D	0.56 ± 0.02	0.61 ± 0.03	0.62 ± 0.03	0.60 ± 0.03	0.61 ± 0.02
	G4D	0.65 ± 0.03	0.70 ± 0.02	0.69 ± 0.03	0.71 ± 0.02	*0.64 ± 0.02*
Weighted_NetEmd	G3D	0.55 ± 0.02	0.60 ± 0.03	0.60 ± 0.03	0.59 ± 0.03	0.60 ± 0.03
	G4D	0.63 ± 0.03	0.68 ± 0.02	0.68 ± 0.03	0.69 ± 0.02	*0.63 ± 0.03*
PCA_NetEmd	50% Variance	0.67 ± 0.03	0.71 ± 0.03	0.72 ± 0.03	0.71 ± 0.02	0.56 ± 0.01
	80% Variance	0.67 ± 0.03	0.69 ± 0.03	0.67 ± 0.02	0.71 ± 0.02	*0.64 ± 0.02*
	90% Variance	0.69 ± 0.03	0.69 ± 0.02	0.69 ± 0.03	0.70 ± 0.02	*0.63 ± 0.02*
	95% Variance	0.65 ± 0.03	0.70 ± 0.02	0.68 ± 0.02	0.71 ± 0.02	0.60 ± 0.01
	99% Variance	0.66 ± 0.03	0.69 ± 0.03	0.68 ± 0.02	0.70 ± 0.02	0.58 ± 0.02
ICA_NetEmd	2 components	0.63 ± 0.03	0.67 ± 0.03	0.68 ± 0.03	0.70 ± 0.03	0.57 ± 0.02
	10 components	0.64 ± 0.02	0.69 ± 0.03	0.70 ± 0.03	0.68 ± 0.02	0.56 ± 0.01
	15 components	0.64 ± 0.02	0.70 ± 0.02	0.68 ± 0.02	0.70 ± 0.02	0.58 ± 0.02
TriadEMD		0.69 ± 0.03	0.62 ± 0.03	0.65 ± 0.03	0.64 ± 0.03	$\mathbf{0.65 \pm 0.02}$
DGCD	13 orbits	0.70 ± 0.03	*0.79 ± 0.03*	*0.77 ± 0.04*	0.69 ± 0.05	0.59 ± 0.04
	129 orbits	$\mathbf{0.84 \pm 0.03}$	$\mathbf{0.80 \pm 0.03}$	$\mathbf{0.78 \pm 0.04}$	0.74 ± 0.05	0.58 ± 0.05
GDA	G3D	0.66 ± 0.04	*0.79 ± 0.03*	0.73 ± 0.03	$\mathbf{0.81 \pm 0.03}$	0.56 ± 0.05
	G4D	0.66 ± 0.06	0.63 ± 0.05	0.62 ± 0.07	0.61 ± 0.06	*0.64 ± 0.06*

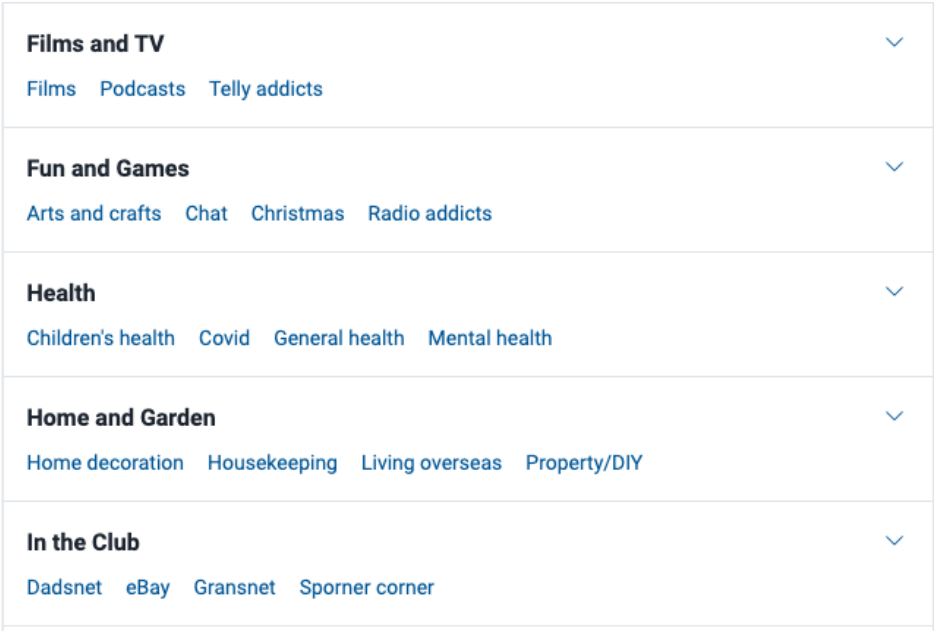
Table B.9: Results for Task 1 in directed networks, for each level of reciprocity. The metric used for evaluation is the mean (and standard error of the mean) of the 16 values for ARI, in each combination of number of nodes, network density and reciprocity. The parameter column indicates: graphlet size used in directed and weighted NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 4); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 4; the number of orbits used by DGCD; graphlet sizes used in GDA. The bolded values are the maximum for each reciprocity level.

Algorithm	Parameter	Synthetic					Real Networks
		Reciprocity					
		0%	25%	50%	75%	100%	
All Orbits	G3D	0.432	0.423	**0.464**	0.387	0.188	0.167
	G4D	0.395	0.409	0.348	0.433	0.439	0.308
Weighted	G3D	0.393	0.361	0.343	0.319	0.311	0.449
	G4D	0.362	0.324	0.319	0.378	0.422	0.373
PCA	50% Variance	0.448	0.274	0.456	0.439	0.429	0.153
	80% Variance	0.425	0.371	0.338	0.396	0.424	0.124
	90% Variance	0.415	0.460	0.350	0.392	0.448	0.192
	95% Variance	0.427	0.468	0.440	**0.475**	0.445	0.151
	99% Variance	0.431	**0.481**	0.394	0.451	0.434	0.159
ICA	2 components	0.451	0.261	0.336	0.403	**0.484**	0.118
	10 components	0.399	0.419	0.397	0.284	0.417	0.282
	15 components	0.418	0.441	0.431	0.438	0.394	0.172
TriadEMD		**0.539**	0.467	0.436	0.392	0.001	0.297
DGCD	13 orbits	0.428	0.445	0.420	0.264	0.154	0.436
	129 orbits	0.432	0.402	0.421	0.386	0.373	0.203
GDA	G3D	0.354	0.255	0.178	0.270	0.000	**0.542**
	G4D	-	-	-	-	0.455	-

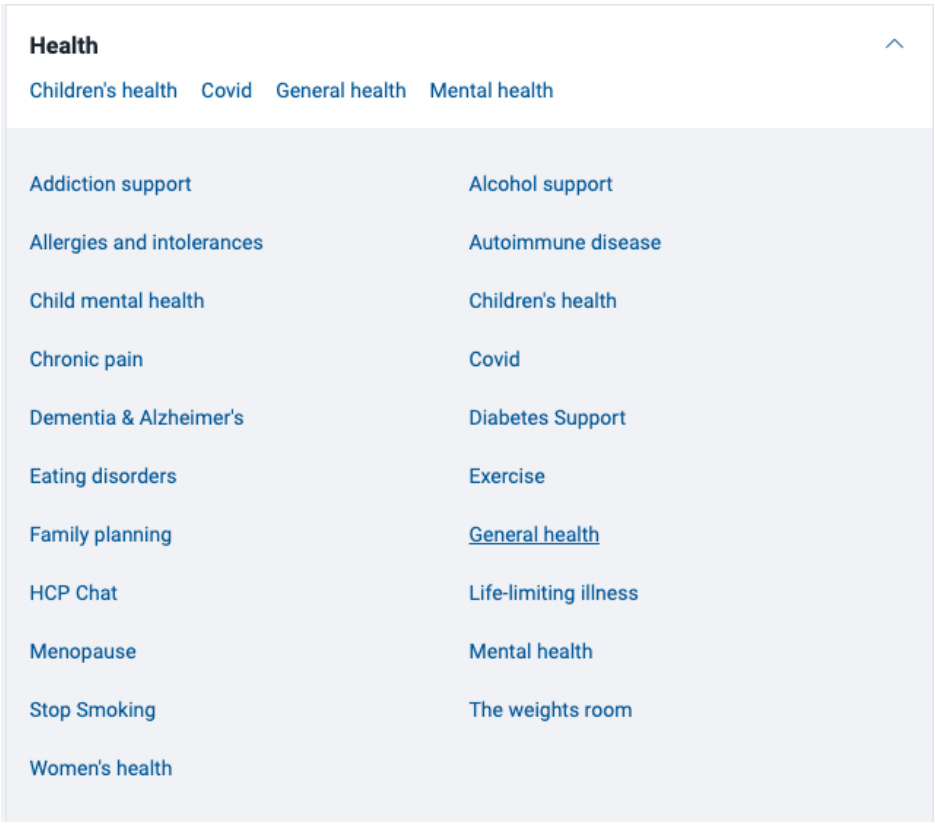
Table B.10: Results for Task 2 in directed networks, for each level of reciprocity, and for the real world networks dataset. The metric used for this task is the sole value of ARI after comparing the 1280 and 1231 networks, respectively. The parameter column indicates: graphlet size used in directed and weighted NetEmd; percentage of variance explained to determine the number of components in *PCA_NetEmd* (using orbits in graphlets of size up to 4); the number of components used in *ICA_NetEmd*, using orbits in graphlets of size up to 4; the number of orbits used by DGCD; graphlet sizes used in GDA. The bolded values are the maximum for each dataset. Note that results for GDA with graphlet size 4 took longer than a week to return results for Task 2 and in the real world networks dataset, at which point we stopped the computation.

Appendix C

Examples of Mumsnet threads



(a) Examples of broad categories of topics within Mumsnet.



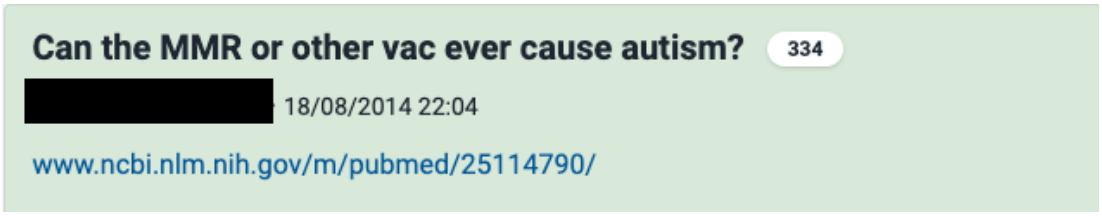
(b) The multiple subcategories of the *Health* topic within Mumsnet. Screenshot was taken after the closure of the vaccination subcategories.

Figure C.1: Organization of the Mumsnet forum, it is divided in broad categories of topics that are expanded into multiple subcategories, each with their own forum.

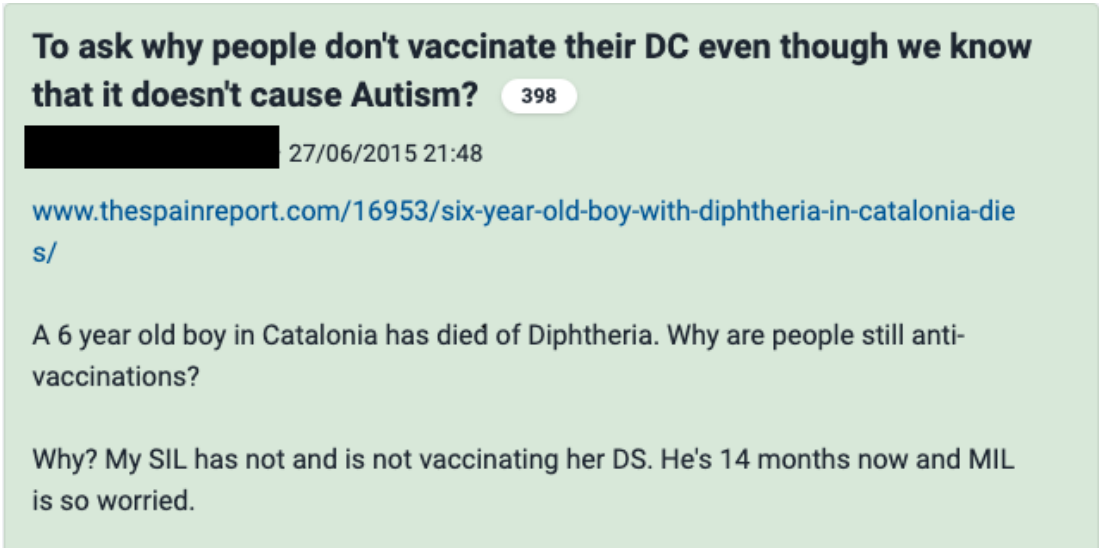
Why has the number of people being vaccinated reduced over the last week?	4	23/02/2021 08:41
I genuinely don't get why people are so offended by small rule breaks at this point	168	23/02/2021 08:39
Your roadmap out of lockdown (light hearted)	6	23/02/2021 08:30
Area stuck on group 5, feeling left behind	4	23/02/2021 08:30
I worry about people's mental health on here if the government fuck it up as usual	53	23/02/2021 07:22
Covid vaccine safe?	81	23/02/2021 06:59

Figure C.2: Examples of threads within the COVID-19 subforum of Mumsnet. In this case, two threads are included in our analysis, the first and last on the list.

Figure C.3: Example of the organization of a thread within Mumsnet. The *original post* is coloured green and the remaining posts are shown sequentially according to the time they were posted in.

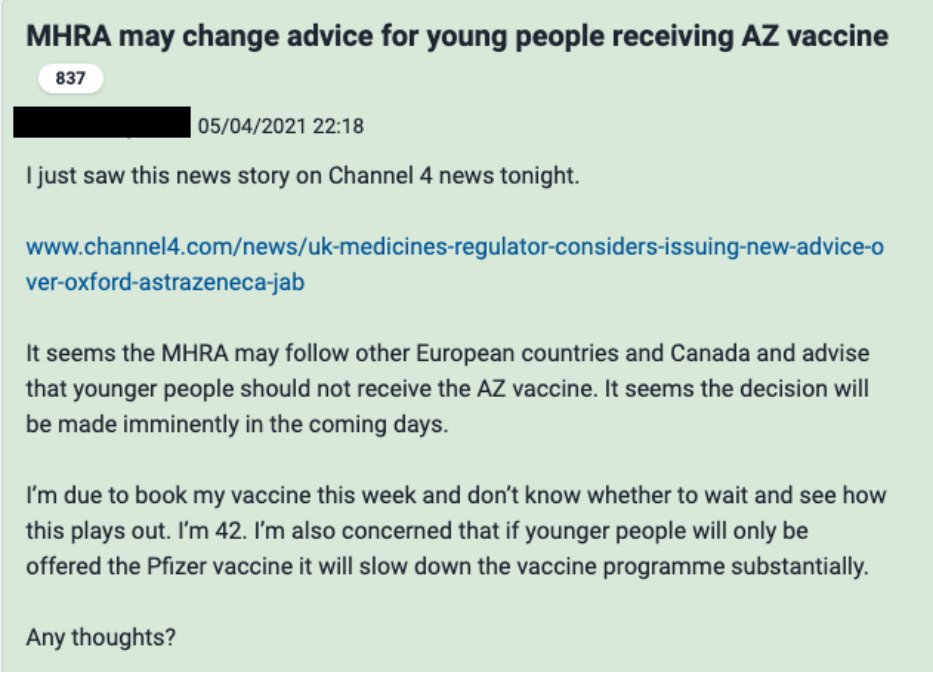


(a) Thread about a paper by Hooker [67] concerning a possible link between the measles mumps rubella (MMR) vaccine and autism, that was later retracted. This thread was posted in the time window spanning from the start of May 2014 to the end of August 2014, where we observe a difference in network structure. This thread was assigned the sentiment of *negative*.

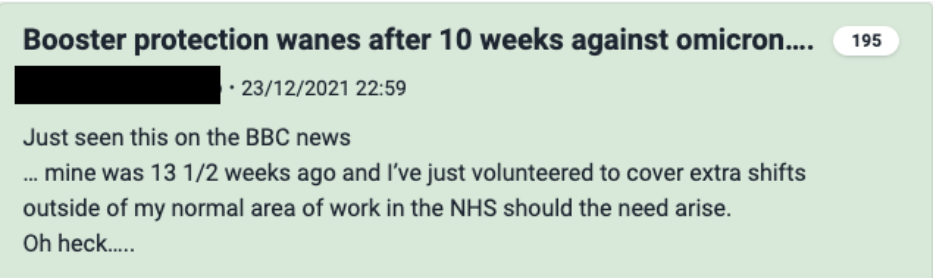


(b) Thread about the death of unvaccinated child due to diphtheria in Spain. This thread was posted in the time window spanning from the start of May 2015 to the end of August 2015, where we observe a difference in network structure. This thread was assigned the sentiment of *positive*.

Figure C.4: Examples of controversial threads in the vaccination forum of Mumsnet that may be connected to the network structure differences we observe in the time frame when these threads were posted.

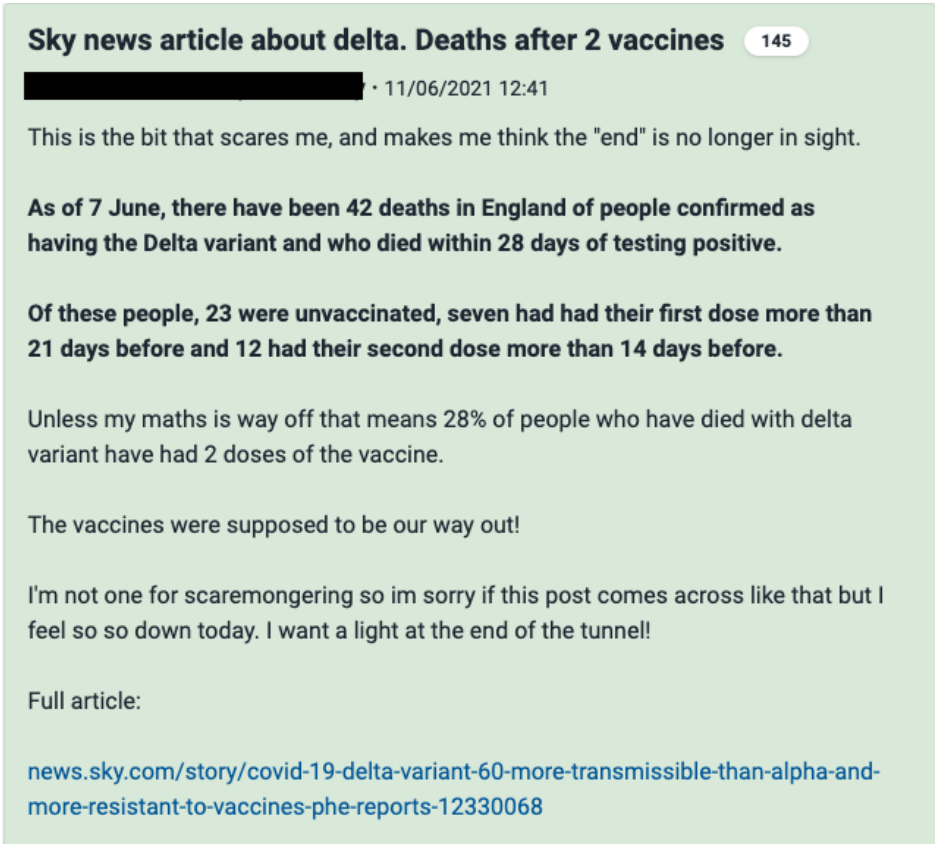


(a) Thread about a link between blood clots and the Astra Zeneca vaccine. This thread was posted in the time window spanning 15 March 2021 to 15 April 2021, where we observe a difference in network structure. This thread was assigned the sentiment of *negative*.



(b) Thread about the booster vaccine in the context of the Omicron variant of COVID-19. This thread was posted in the time window spanning 1 December 2021 to 1 January 2022, where we observe a difference in network structure. This thread was assigned the sentiment of *neutral*.

Figure C.5: Examples of discussions about vaccination in the COVID-19 forum of Mumsnet with representative topics that may be connected to the network structure differences we observe in the time frame when these threads were posted.



(c) Thread about the efficacy of vaccination against the Delta variant of COVID-19. This thread was posted in the time window spanning 1 June 2021 to 1 July 2021, where we observe a difference in network structure. This thread was assigned the sentiment of *negative*.

Figure C.5: Continued - Examples of discussions about vaccination in the COVID-19 forum of Mumsnet with representative topics that may be connected to the network structure differences we observe in the time frame when these threads were posted.

Appendix D

Mumsnet - supporting figures

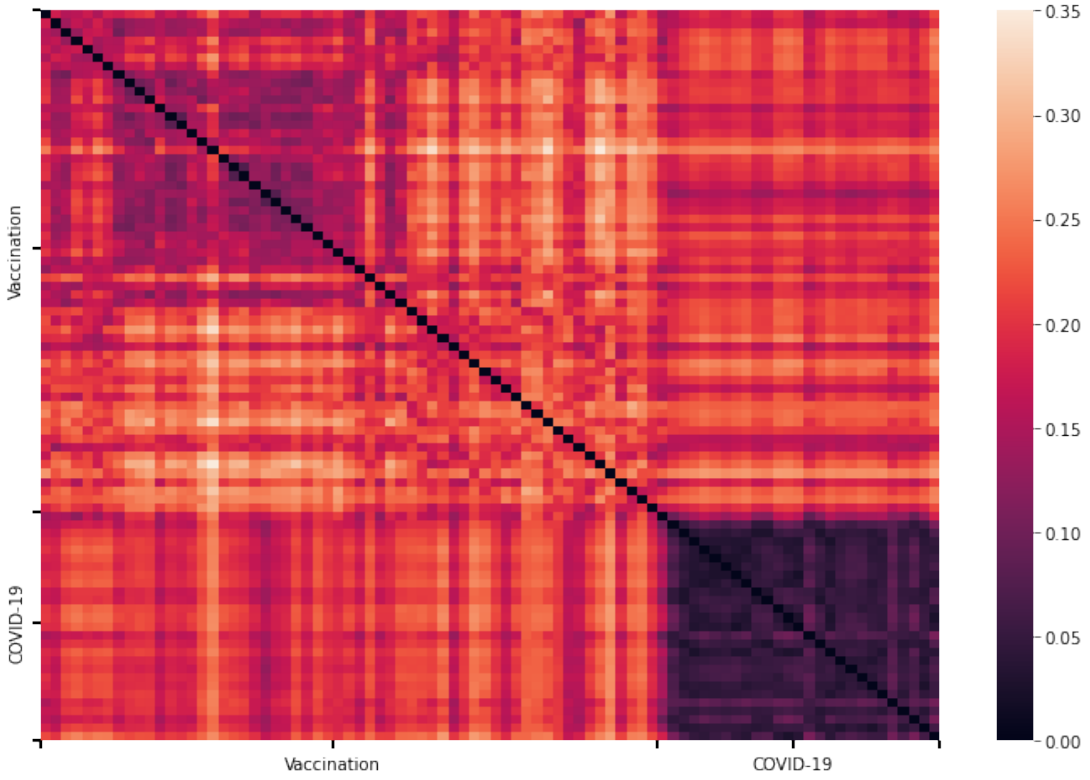


Figure D.1: Heat map of network distances between all networks in both the Mumsnet vaccination and COVID-19 dataset. Distances are calculated using *PCA_NetEmd* with size 4 graphlets and 90% explained variance.

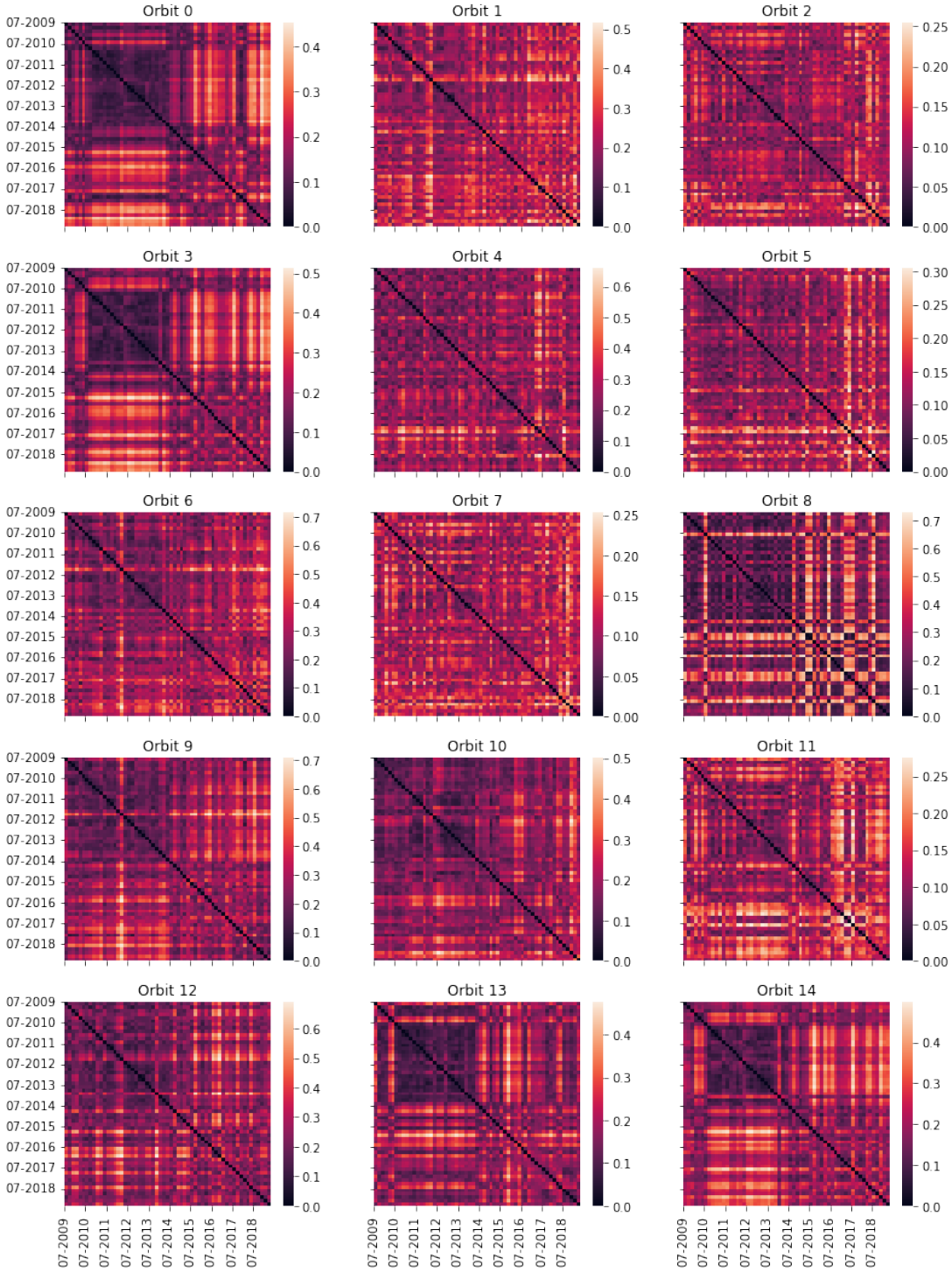


Figure D.2: Heat map of network distances between each pair of networks in the Mumsnet vaccination dataset, for each orbit in graphlets of size up to 4. Distances are calculated using *PCA_NetEmd* with size 4 graphlets and 90% explained variance.

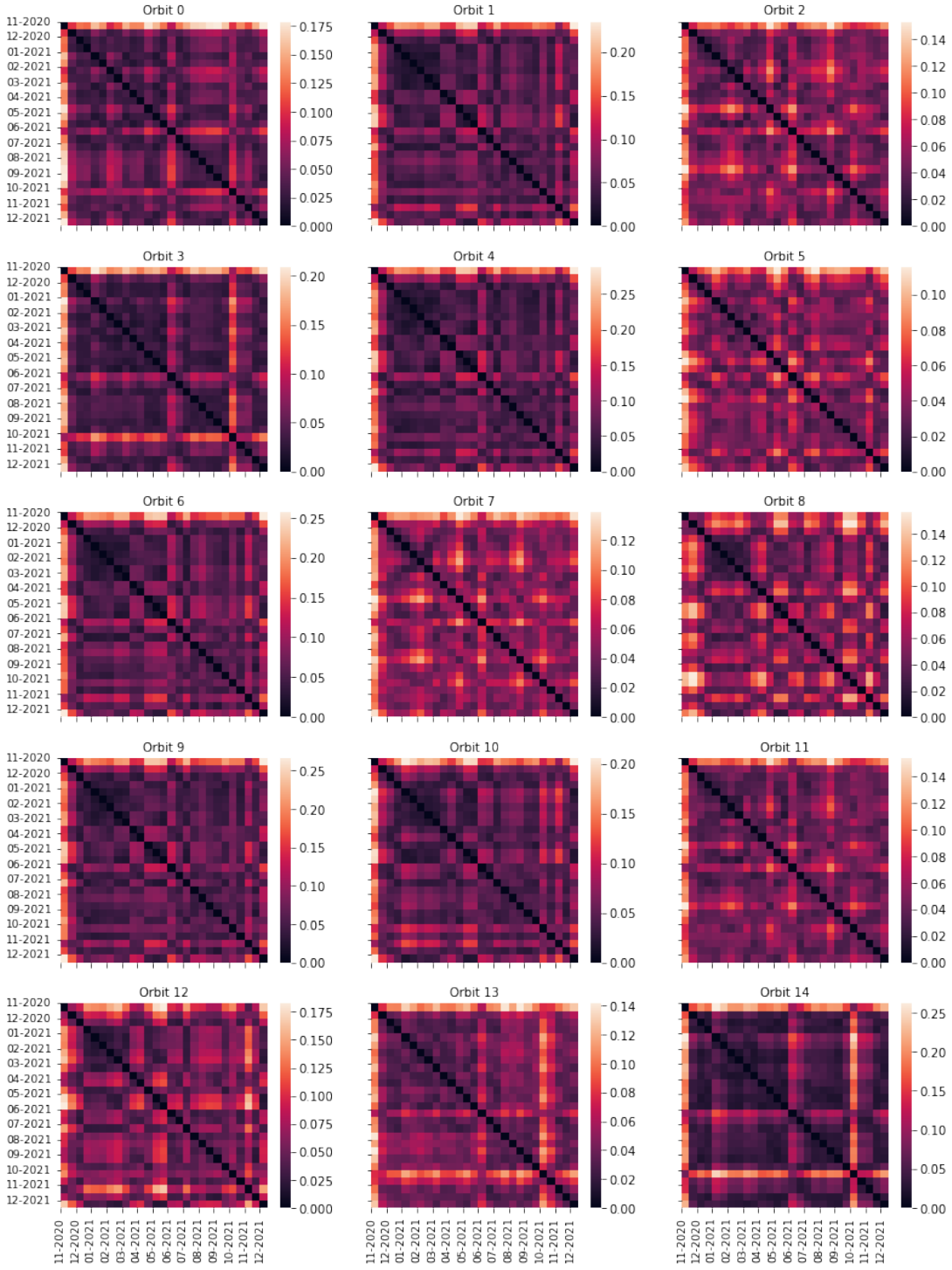


Figure D.3: Heat map of network distances between each pair of networks in the Mumsnet COVID-19 dataset, for each orbit in graphlets of size up to 4. Distances are calculated using *PCA_NetEmd* with size 4 graphlets and 90% explained variance.

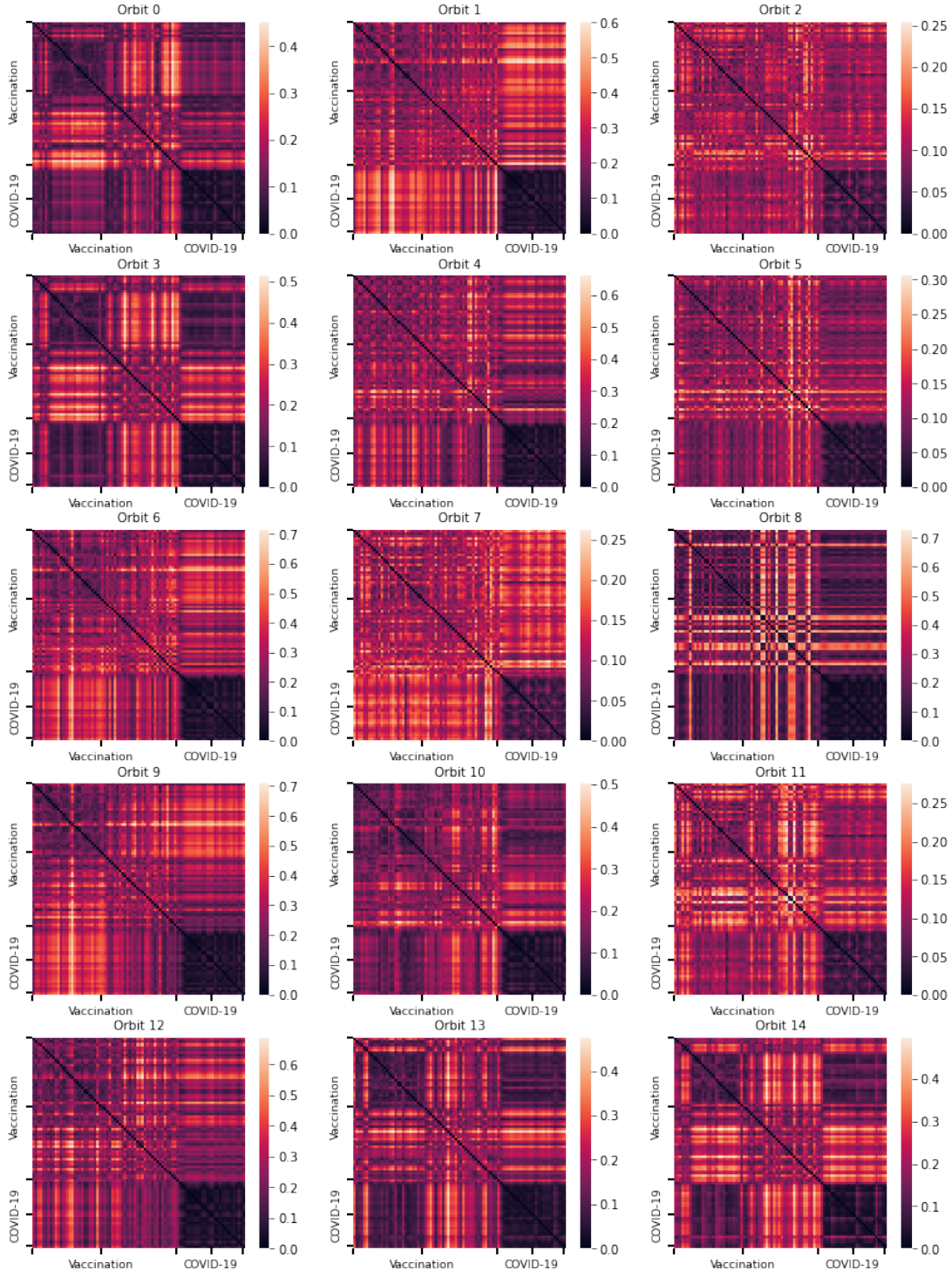
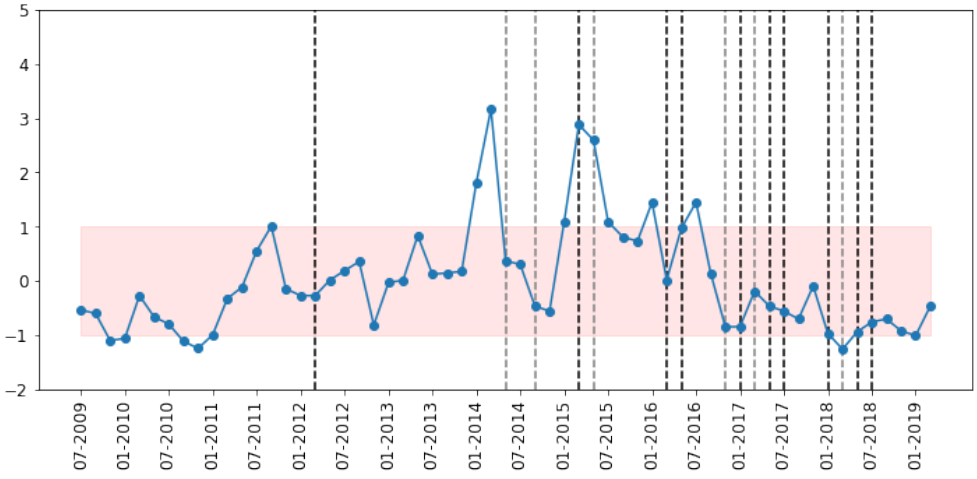
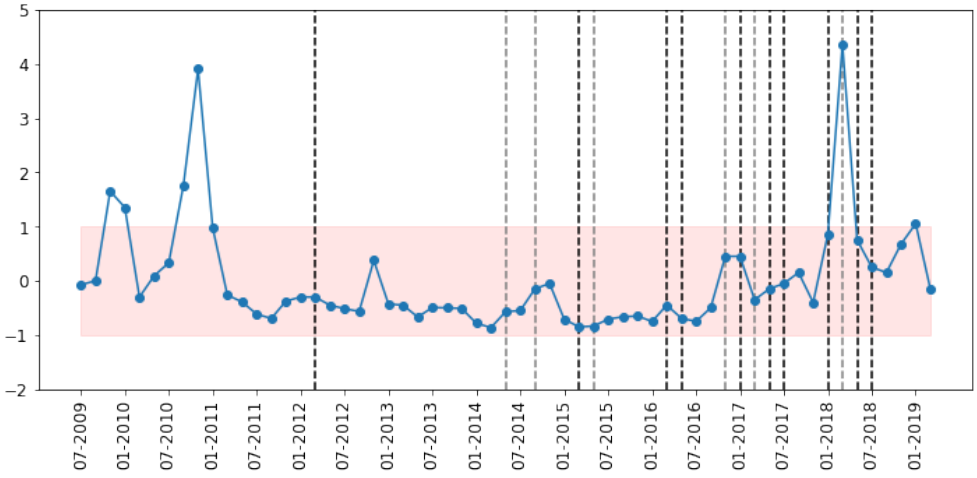


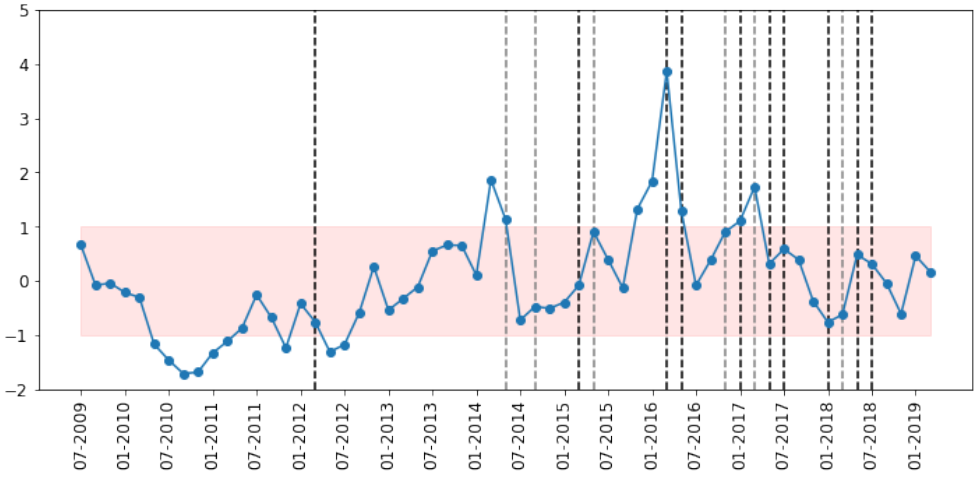
Figure D.4: Heat map of network distances between all networks in both the Mumsnet vaccination and COVID-19 dataset, for each orbit in graphlets of size up to 4. Distances are calculated using *PCA_NetEmd* with size 4 graphlets and 90% explained variance.



(a) Proportion of positive to negative threads.

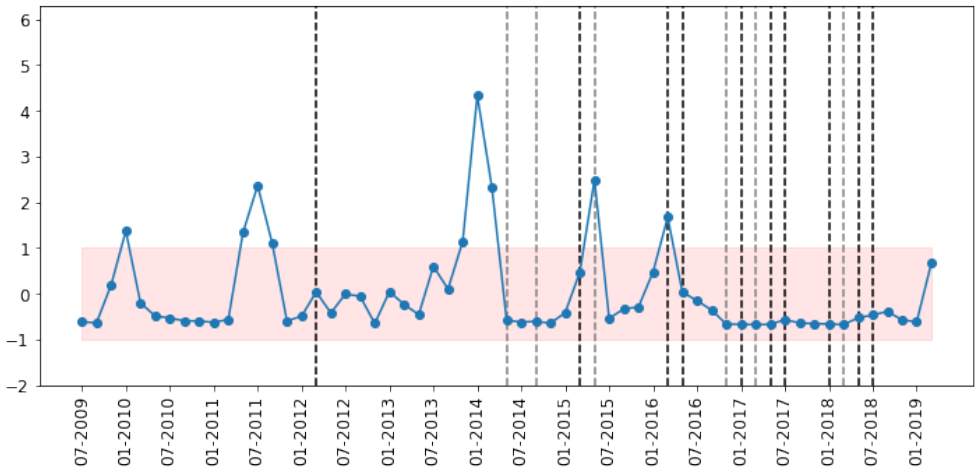


(b) Proportion of negative to positive threads.

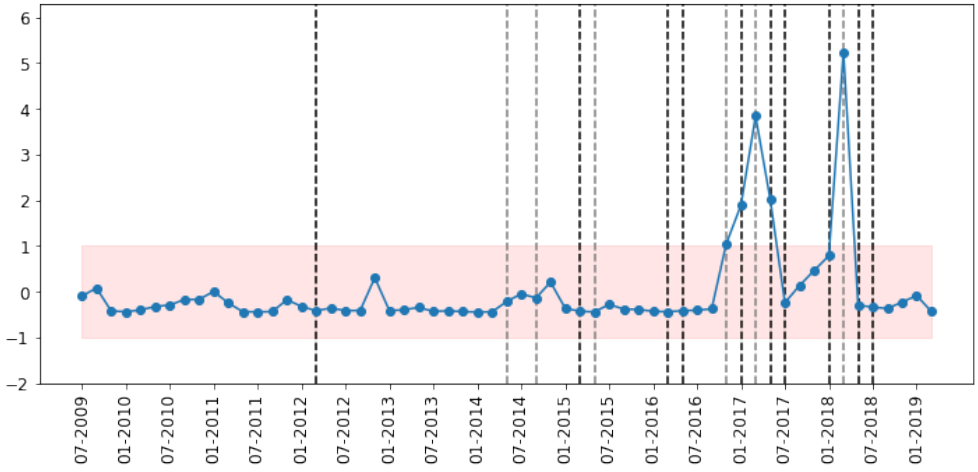


(c) Proportion of neutral to positive and negative threads.

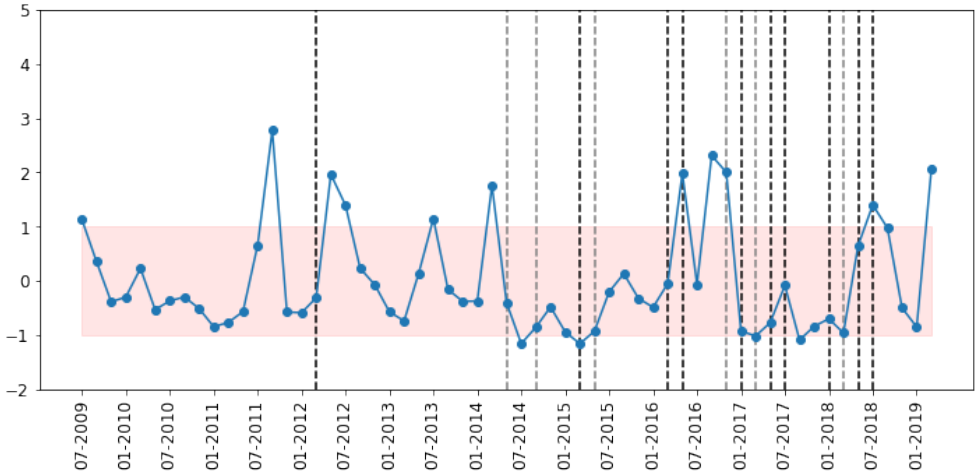
Figure D.5: Vaccination subforum - threads of each sentiment.



(a) Proportion of posts in positive threads to posts in negative threads.

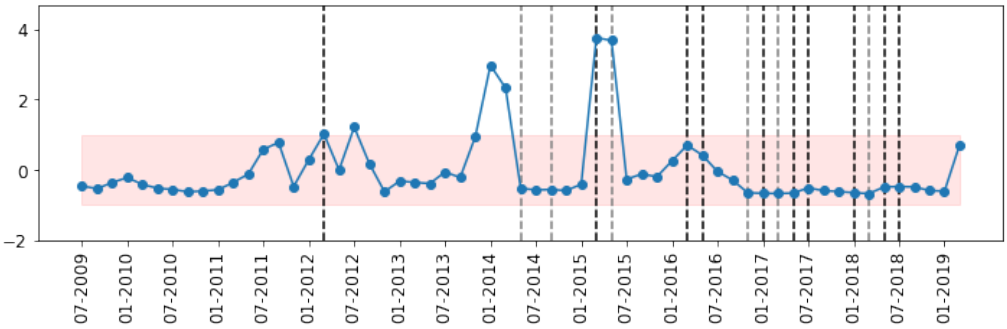


(b) Proportion of posts in negative threads to posts in positive threads.

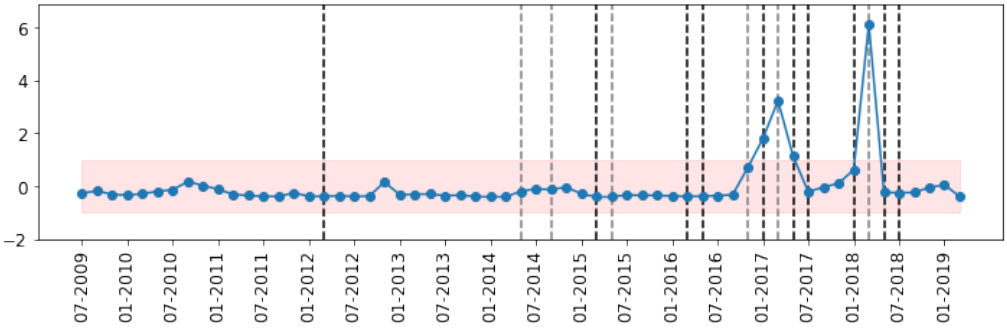


(c) Proportion of posts in neutral threads to posts in positive and negative threads.

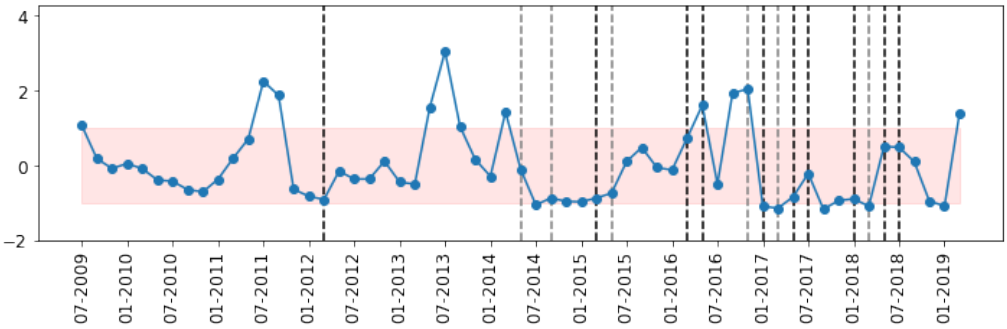
Figure D.6: Vaccination subforum - posts in threads of each sentiment.



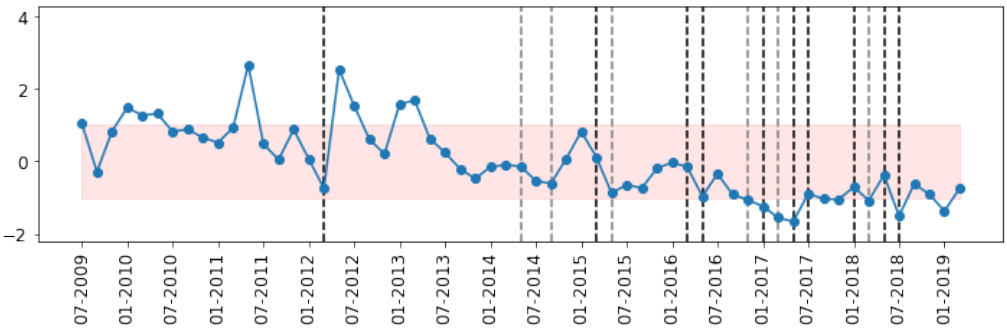
(a) Proportion of users in positive threads to users in negative threads.



(b) Proportion of users in negative threads to users in positive threads.

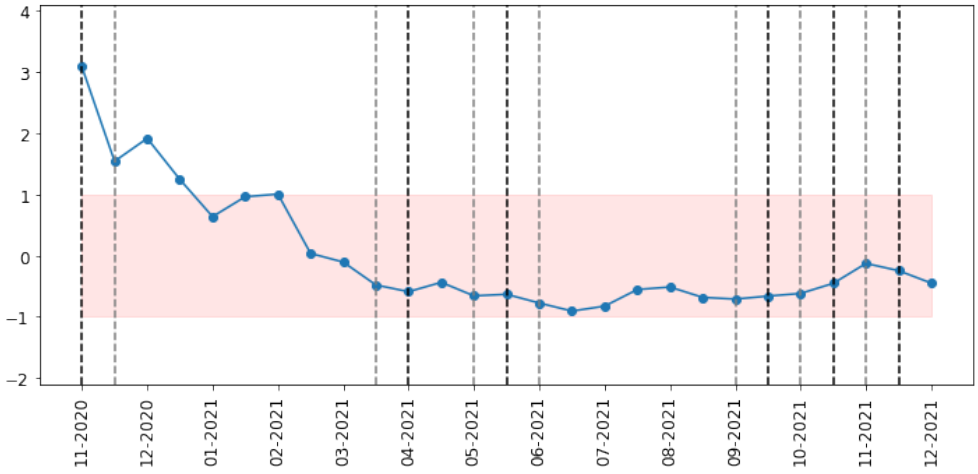


(c) Proportion of users in neutral threads to users in positive and negative threads.

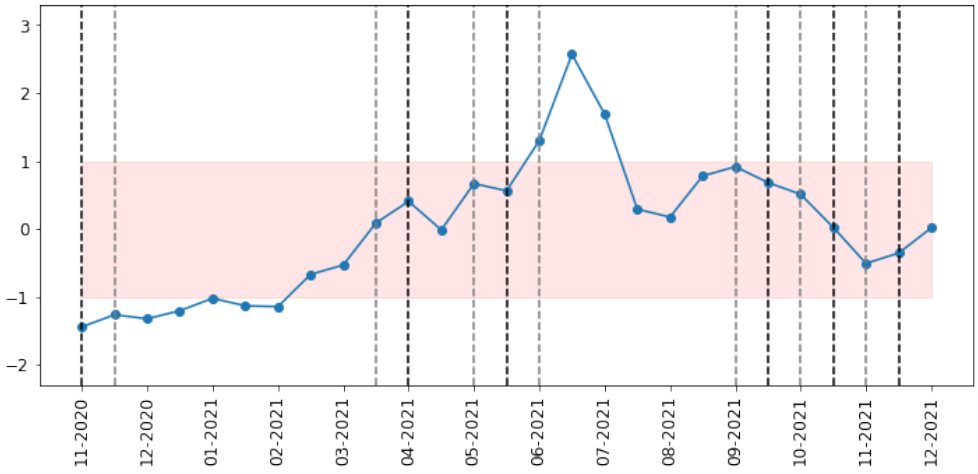


(d) Proportion of users that post in multiple threads with different sentiments to users only in neutral, positive or negative threads.

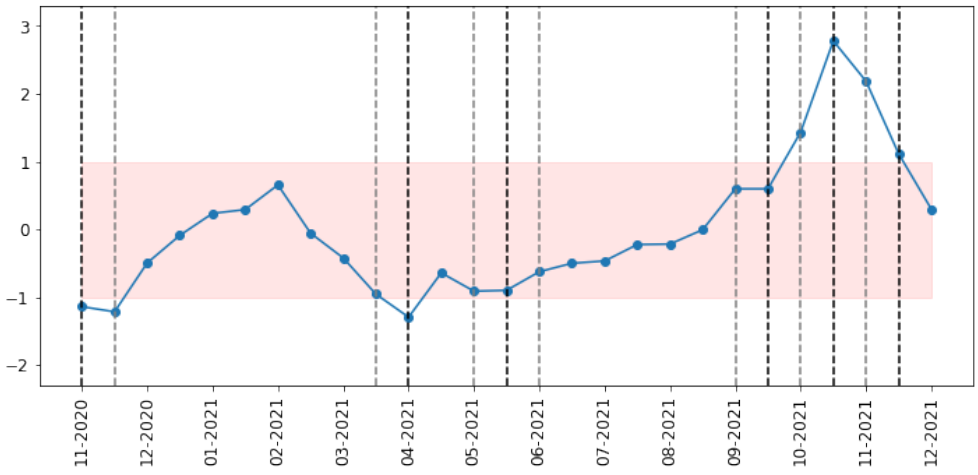
Figure D.7: Vaccination subforum - users in threads of each sentiment.



(a) Proportion of positive to negative threads.

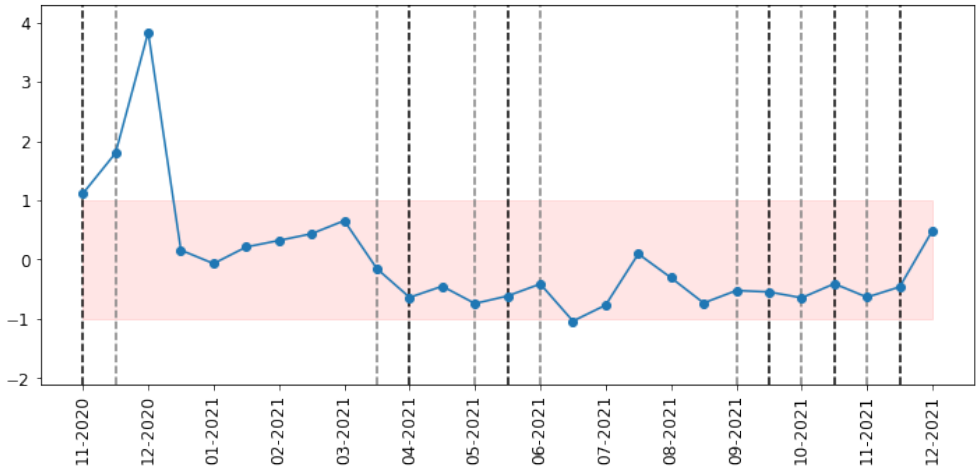


(b) Proportion of negative to positive threads.

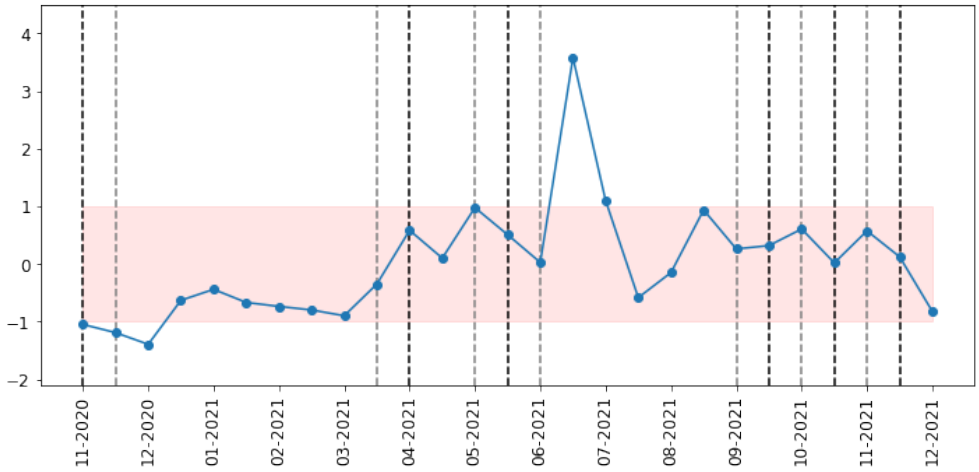


(c) Proportion of neutral to positive and negative threads.

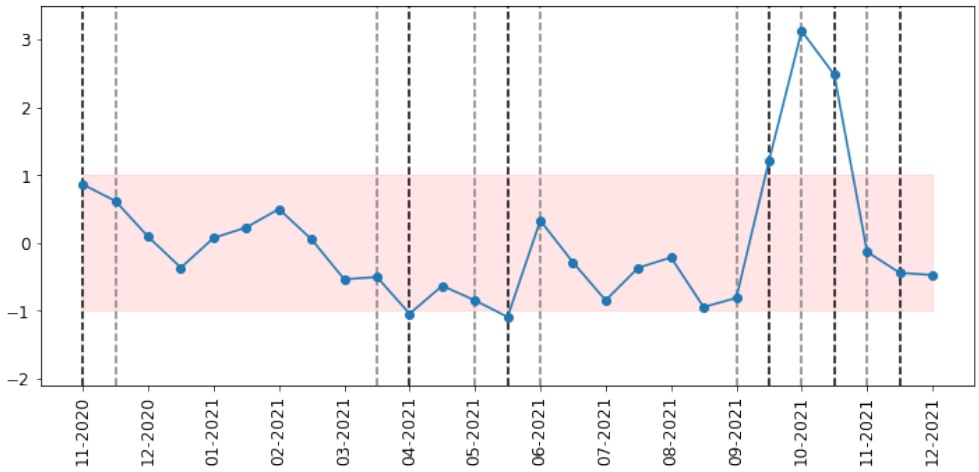
Figure D.8: COVID-19 subforum - threads of each sentiment.



(a) Proportion of posts in positive threads to posts in negative threads.

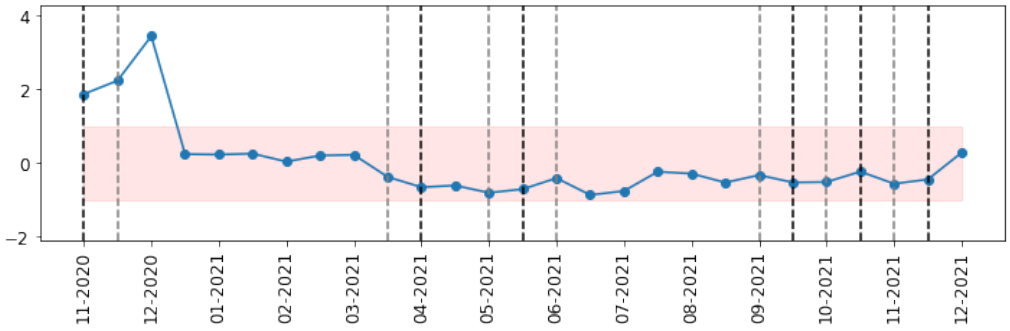


(b) Proportion of posts in negative threads to posts in positive threads.

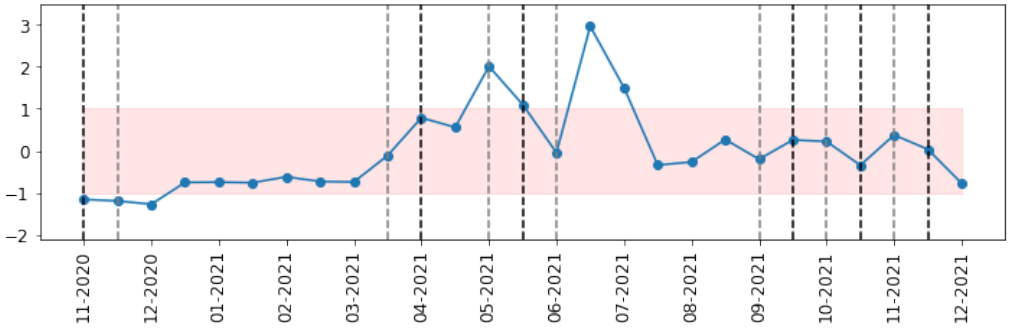


(c) Proportion of posts in neutral threads to posts in positive and negative threads.

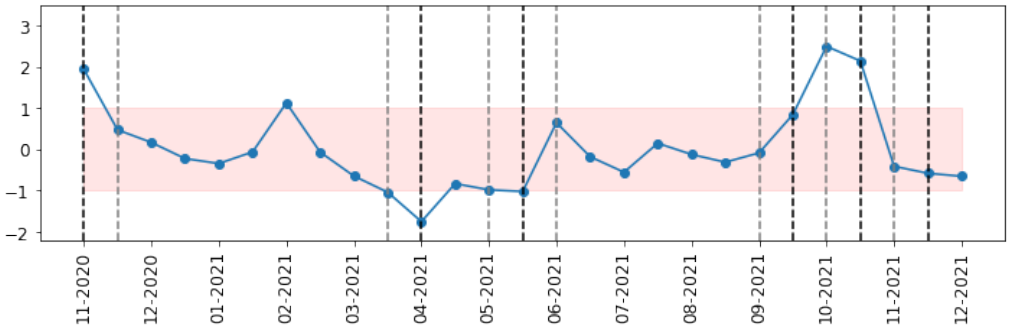
Figure D.9: COVID-19 subforum - posts in threads of each sentiment.



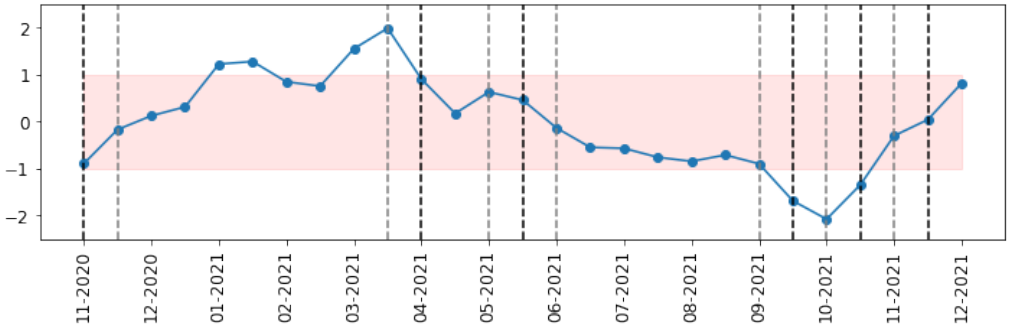
(a) Proportion of users in positive threads to users in negative threads.



(b) Proportion of users in negative threads to users in positive threads.

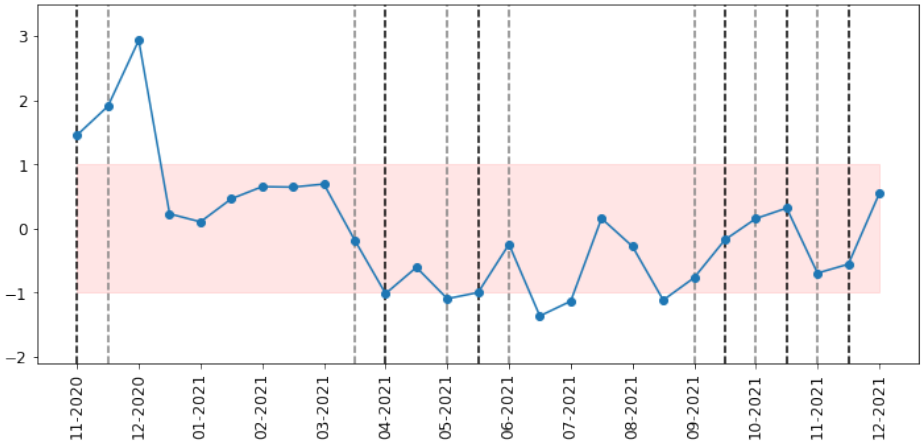


(c) Proportion of users in neutral threads to users in positive and negative threads.

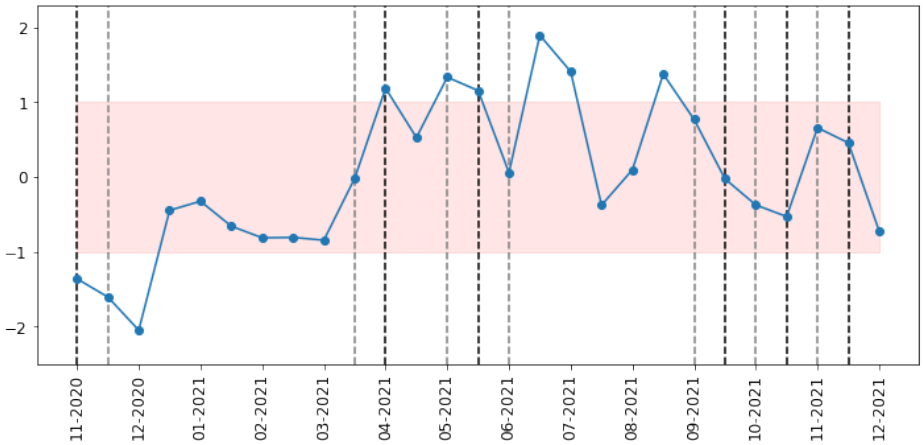


(d) Proportion of users that post in multiple threads with different sentiments to users only in neutral, positive or negative threads.

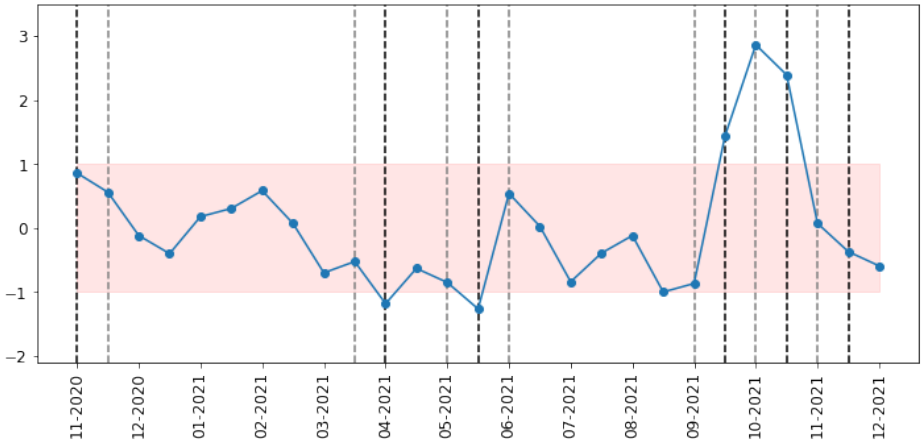
Figure D.10: COVID-19 subforum - users in threads of each sentiment.



(a) Proportion of positive to negative posts.



(b) Proportion of negative to positive posts.



(c) Proportion of neutral to positive and negative posts.

Figure D.11: COVID-19 subforum - posts of each sentiment, inferred from the sampled subset of data.

Appendix E

Mumsnet - alternative data granularity

In Section 4.4.1, we claim that our conclusions should not depend on the choice of window span and jump. In this appendix, we show how using different values for these parameters leads to similar final results.

E.1 Vaccination forum

The choice of window span and jump for the vaccination dataset was mostly supported on two observations: that a 2 month window jump leads to a good trade-off between enough variability in consecutive data slices and not too much variability that differences become meaningless; and, upon fixing the window jump, the window span follows naturally under the assumption that consecutive data slices sharing more than 50% of date range leads to consecutive data slices that are too similar. In this section, we challenge this assumption by keeping a window jump of two months and a window span of 6 months, showing the impact of consecutive data slices sharing two thirds of the date range.

Upon splitting the vaccination dataset in data slices starting on the first day of each odd-numbered month with data spanning 6 months, we obtain 58 networks, whose summary statistics are shown in Table E.1. The network comparison of these 58 networks is illustrated by the heatmap in Figure E.1. The Z-scored proportion of positive to negative, negative to positive and neutral threads, posts and users are shown in Figure E.2, E.3 and E.4.

Using the previous parameters for data granularity, we observed 16 differences

		Vaccine	COVID-19
# Networks		58	59
Number of Nodes	Min.	189	906
	Median	339	2149
	Max	579	4231
Number of Edges	Min.	514	6639
	Median	1462	37719
	Max	7070	117442
Density	Min.	1.31×10^{-2}	9.57×10^{-3}
	Median	2.74×10^{-2}	1.75×10^{-2}
	Max	4.32×10^{-2}	3.51×10^{-2}
Average Degree	Min.	5.1	14.5
	Median	9.2	36.6
	Max	24.7	70.3
Clustering Coefficient	Min.	0.45	0.58
	Median	0.72	0.77
	Max	0.92	0.87

Table E.1: Summary statistics for the datasets of networks from the vaccination and COVID-19 datasets with alternative choices of window span and jump. The values shown for the vaccination dataset correspond to a span of 6 months and jump of 2 months. For the COVID-19 dataset, the values of these two parameters are 2 weeks and 1 week, respectively.

in network structure against 20 time windows with significant change in sentiment. With the combination of window span and jump we show in this section, we obtain 18 differences in network structure, as we also consider differences from index i to $i+3$ (marked in the figures as a light gray dotted line) due to the necessity of comparing networks with no overlapping data. On the other hand, we recognize a decrease in the number of significant changes in sentiment to 18.

The change in data granularity further accentuates the disparity of network structure before and after 2015. Using either set of parameters, it is clear that the network prior to 2015 are all very similar to each other and after 2015 there is a constant difference in the structure of temporally contiguous networks, with significant differences detected in the majority of time slices. As no significant differences were detected prior to 2015, there are changes in sentiment in 2011 and 2012 that are not reflected in differences of network structure, which were captured when using the shorted window span. On the other hand, with the shorter window span, we saw no changes in sentiment around September 2014, but using a longer window span shows that between the windows starting in July 2014 and March 2015, there is an underrepresentation of posts and users in neutral threads (Figures E.3c and E.4c). The remaining conclusions

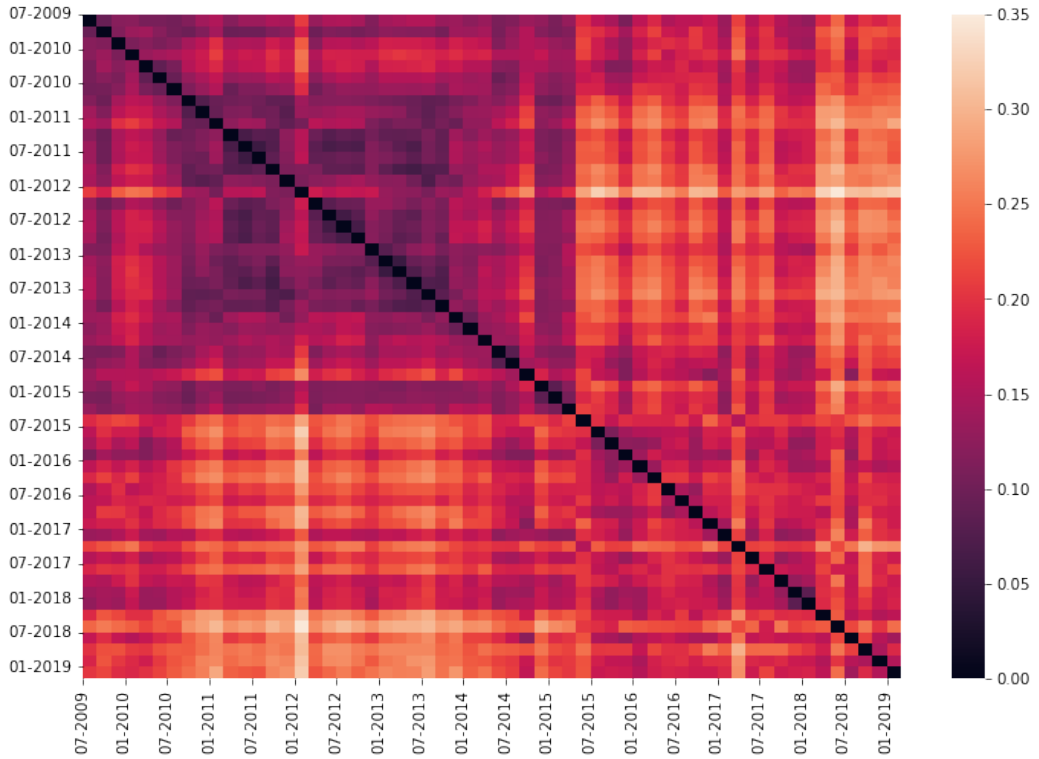


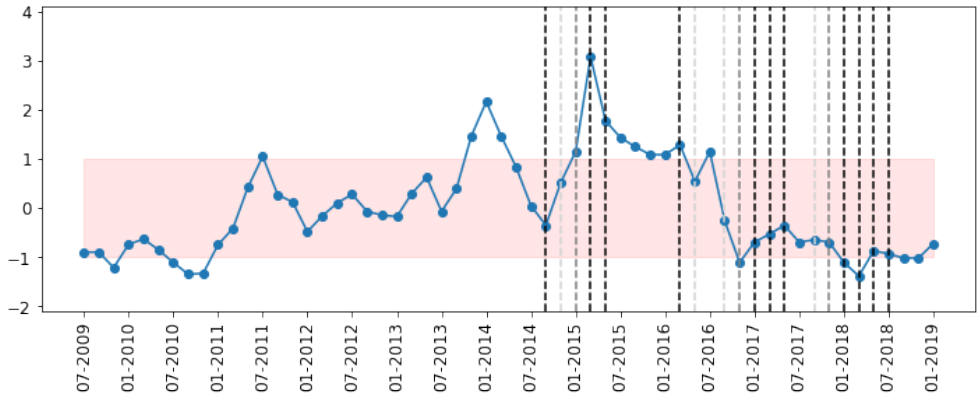
Figure E.1: Heat map of network distances between each pair of networks in the Mumsnet vaccination dataset, using an alternative combination of window span and jump (6 months and 2 months). Distances are calculated using *PCA_NetEmd* with size 4 graphlets and 90% explained variance.

using a shorter window span match those using the alternative longer window span.

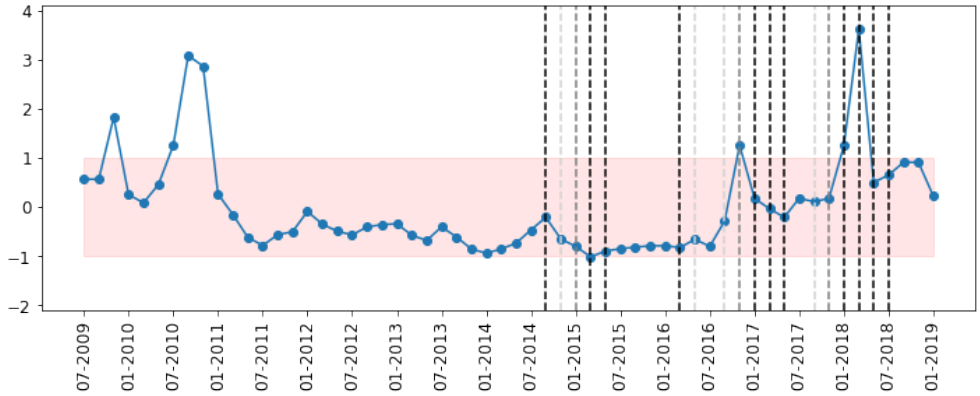
E.2 COVID-19 forum

Unlike the vaccination forum, in the COVID-19 forum we identified two combinations of window span and jump that equally fit the criteria for this dataset. Our choice landed on the combination of one month span with a two week jump, as we argued that the monthly span fit the timescale of policy changes better than the alternative we present in this section, two week span with a one week jump.

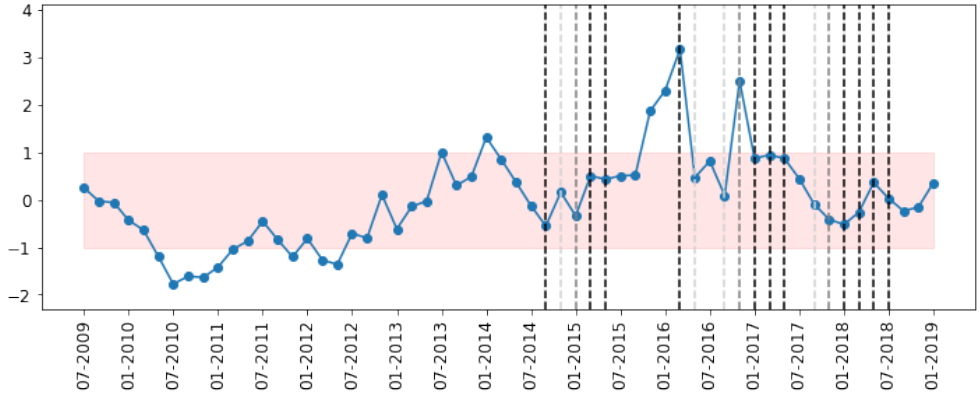
By splitting the COVID-19 dataset in windows spanning two weeks for every week in the dataset, we obtain 59 data slices, whose summary statistics are shown in Table E.1. The network comparison of these 59 networks is illustrated by the heatmap in Figure E.5. The Z-scored proportion of positive to negative, negative to positive and neutral threads, posts and users are shown in Figure E.6, E.7 and E.8.



(a) Proportion of positive to negative threads.

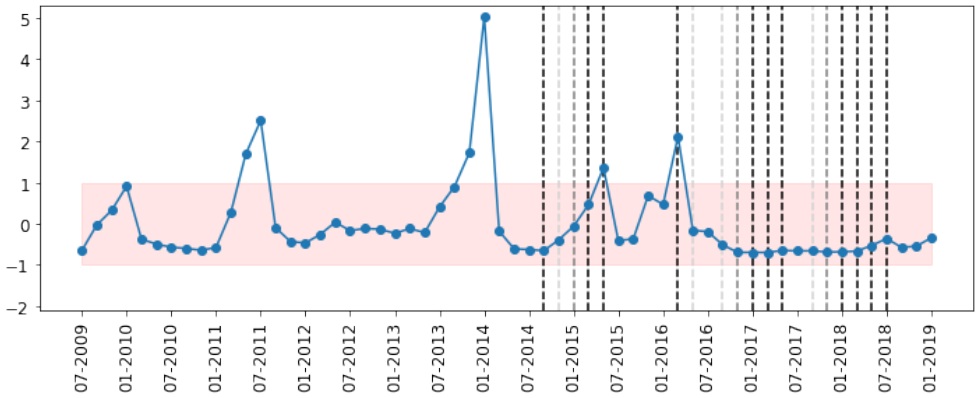


(b) Proportion of negative to positive threads.

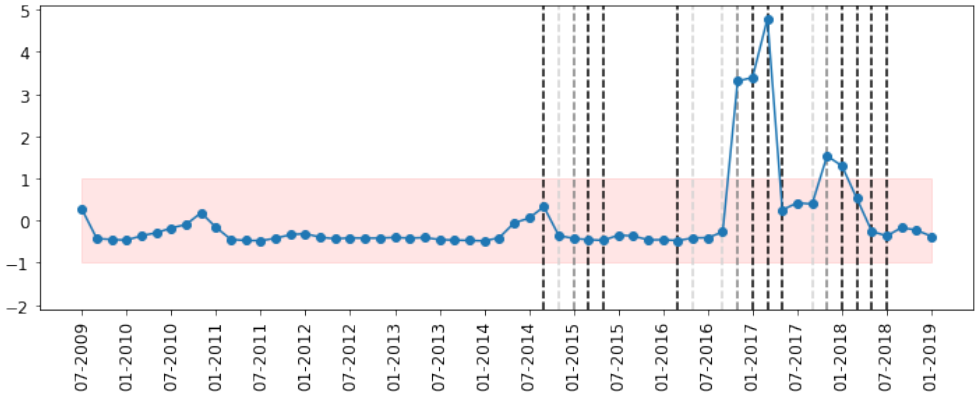


(c) Proportion of neutral to positive and negative threads.

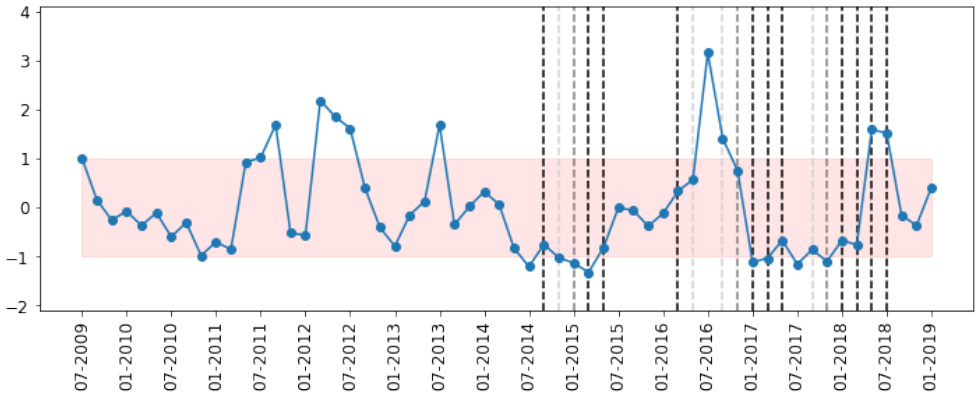
Figure E.2: Vaccination forum - threads of each sentiment - alternative data granularity.



(a) Proportion of posts in positive threads to posts in negative threads.

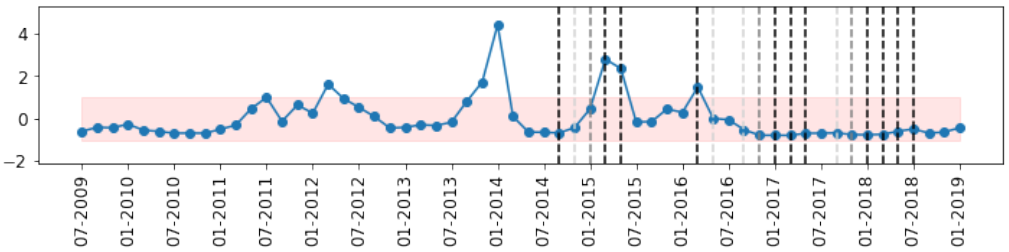


(b) Proportion of posts in negative threads to posts in positive threads.

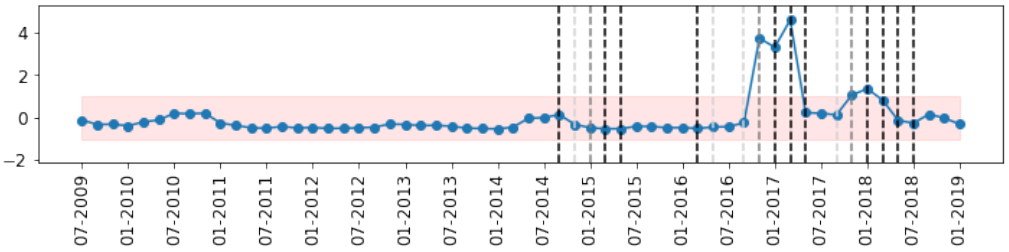


(c) Proportion of posts in neutral threads to posts in positive and negative threads.

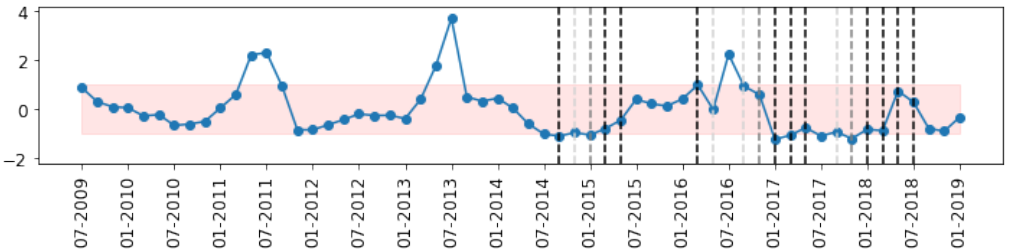
Figure E.3: Vaccination forum - posts in threads of each sentiment - alternative data granularity.



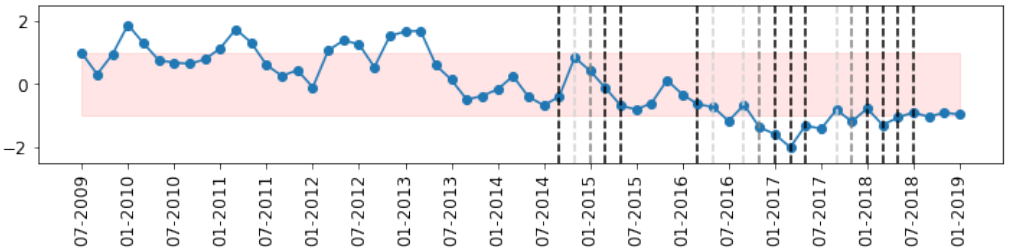
(a) Proportion of users in positive threads to users in negative threads.



(b) Proportion of users in negative threads to users in positive threads.



(c) Proportion of users in neutral threads to users in positive and negative threads.



(d) Proportion of users that post in multiple threads with different sentiments to users only in neutral, positive or negative threads.

Figure E.4: Vaccination forum - users in threads of each sentiment - alternative data granularity.

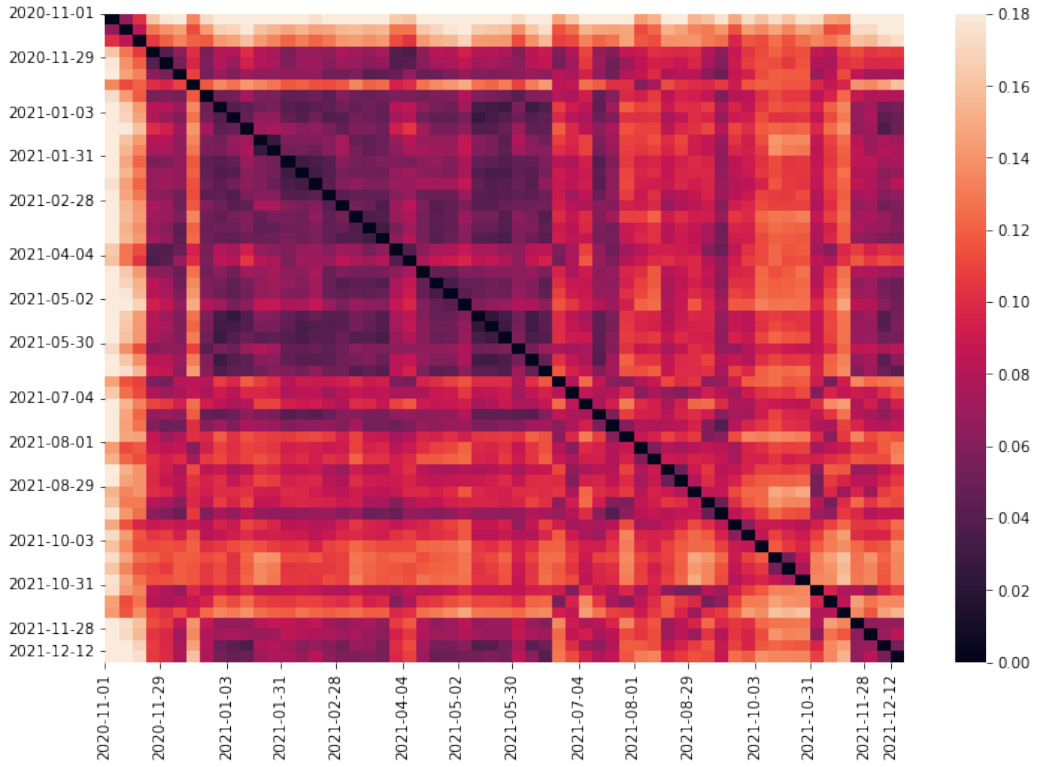
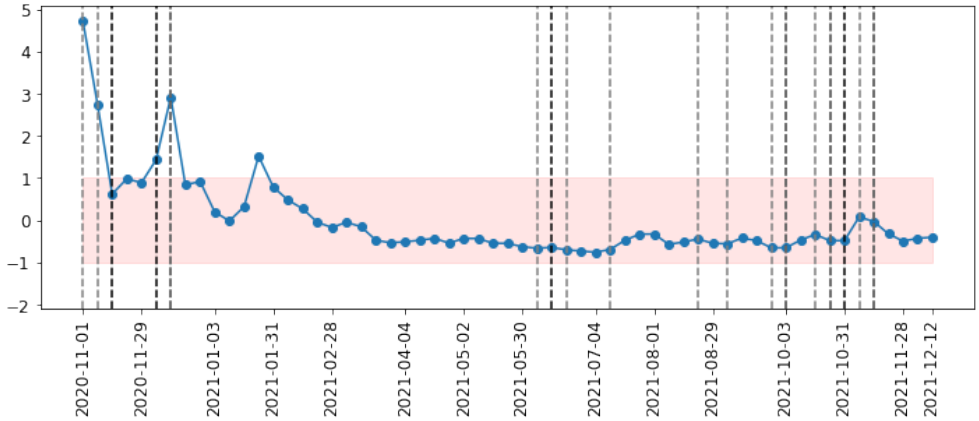
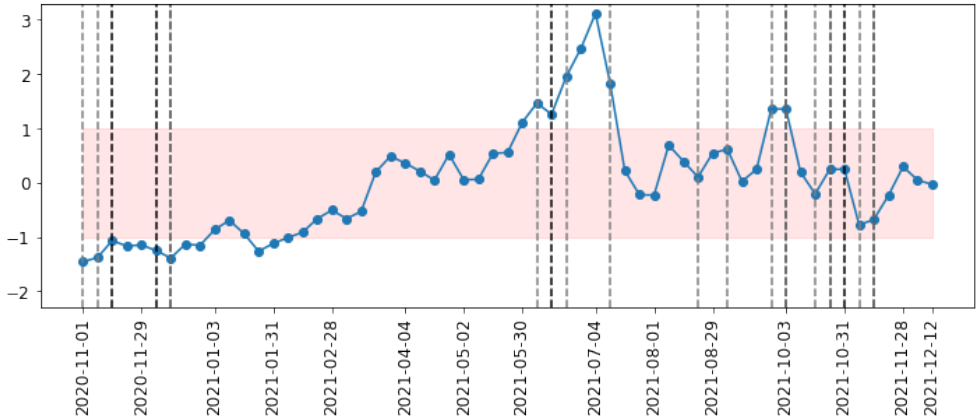


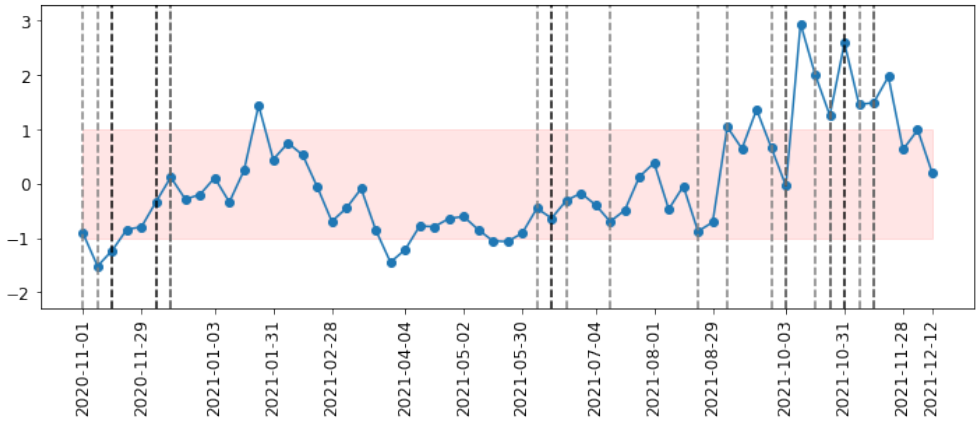
Figure E.5: Heat map of network distances between each pair of networks in the Mumsnet COVID-19 dataset, using an alternative combination of window span and jump (2 weeks and 1 week). Distances are calculated using *PCA_NetEmd* with size 4 graphlets and 90% explained variance.



(a) Proportion of positive to negative threads.

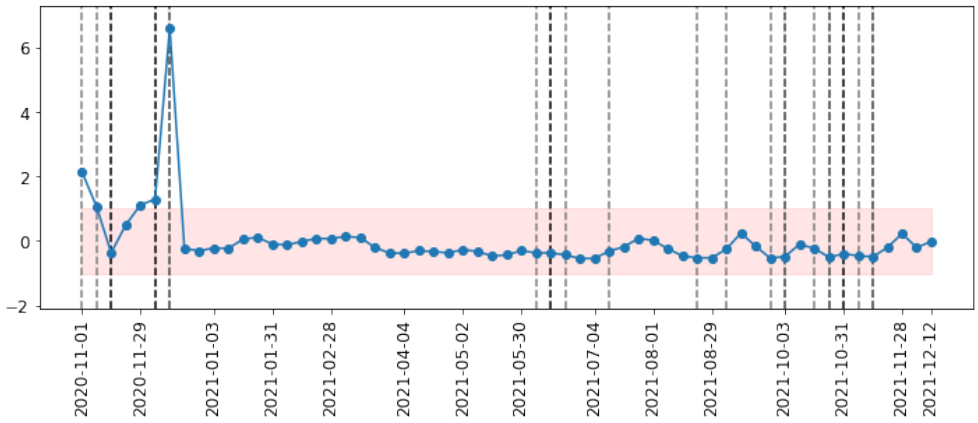


(b) Proportion of negative to positive threads.

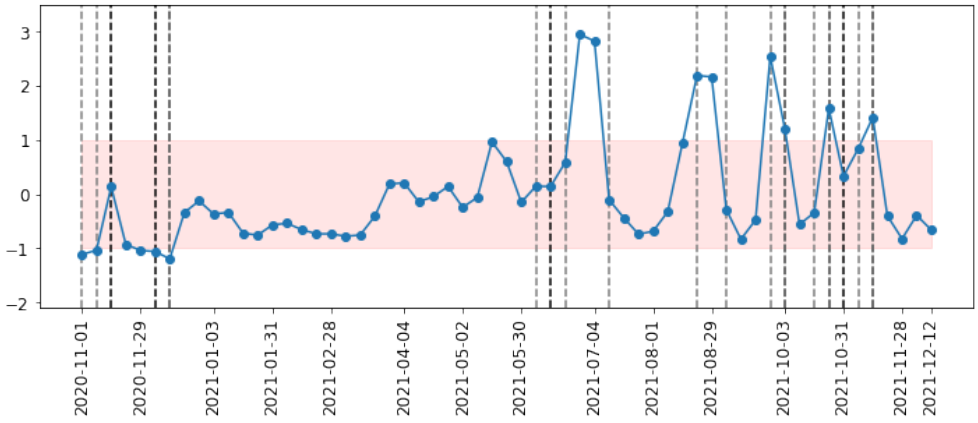


(c) Proportion of neutral to positive and negative threads.

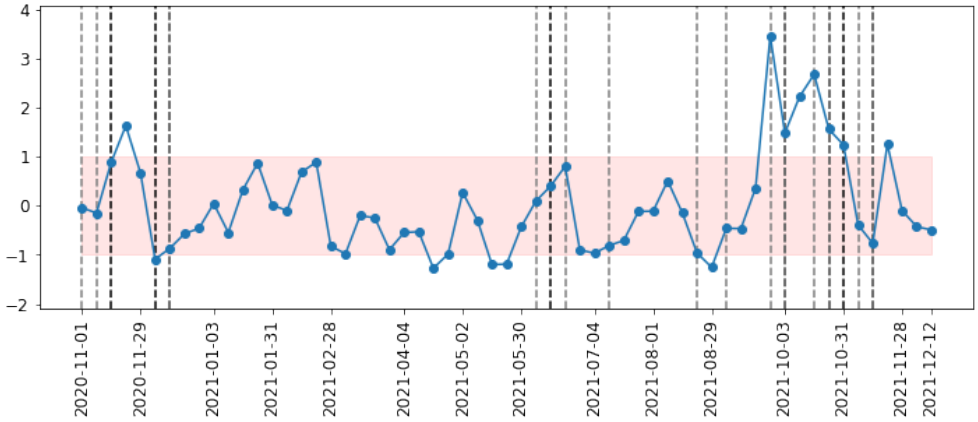
Figure E.6: COVID-19 forum - threads of each sentiment - alternative data granularity.



(a) Proportion of posts in positive threads to posts in negative threads.

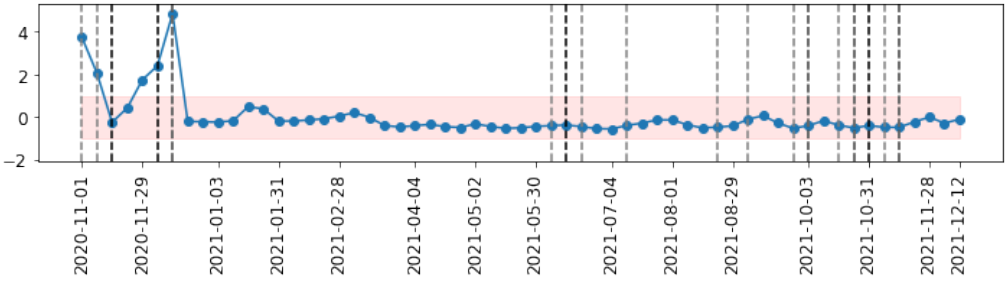


(b) Proportion of posts in negative threads to posts in positive threads.

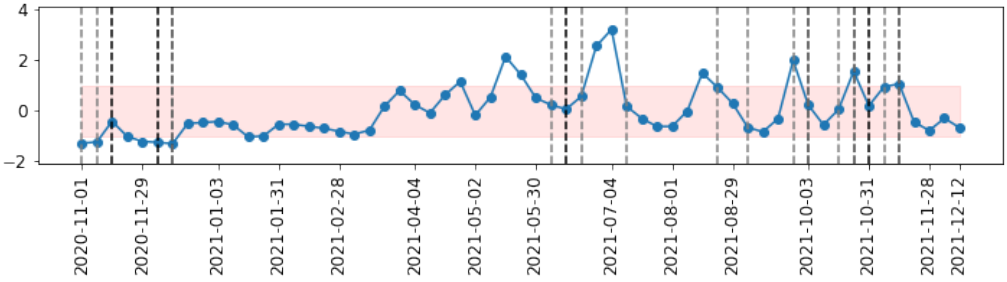


(c) Proportion of posts in neutral threads to posts in positive and negative threads.

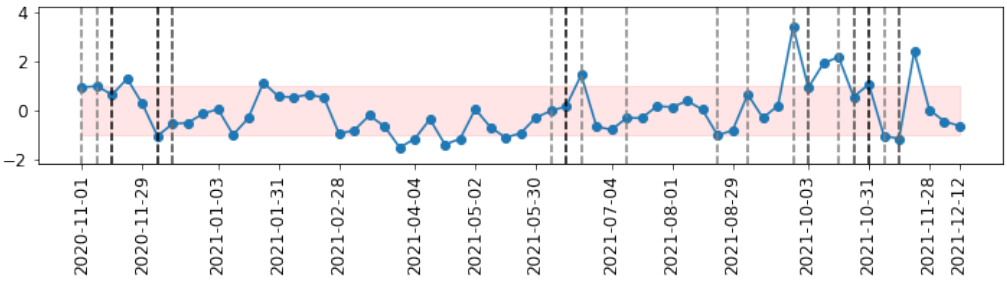
Figure E.7: COVID-19 forum - posts in threads of each sentiment - alternative data granularity.



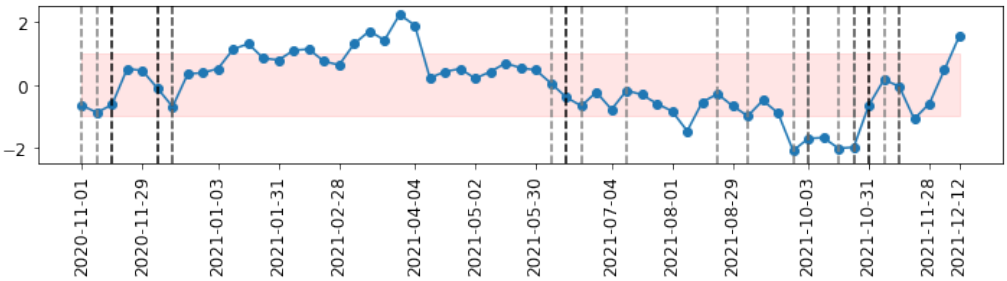
(a) Proportion of users in positive threads to users in negative threads.



(b) Proportion of users in negative threads to users in positive threads.



(c) Proportion of users in neutral threads to users in positive and negative threads.



(d) Proportion of users that post in multiple threads with different sentiments to users only in neutral, positive or negative threads.

Figure E.8: COVID-19 subforum - users in threads of each sentiment - alternative data granularity.

Applying the same methodology to identify differences in network structure yields four groups of differences, three of which that match differences in network structure using the previous combination of window jump and span. Using new values for these parameters, differences emerge in the networks at the end of August, particularly between the slices starting on 22 of August of 2021 and 5 of September 2021. On the other hand, there are no differences in network structure highlighted in March and April of 2021, as observed with the 1 month span and 2 week jump parameter combination.

These groups of differences in network structure are related to changes in sentiment in the following way:

- The first group, spanning networks starting in 1 of November 2020 until 13 of December 2020, are related to an overrepresentation of positive sentiment (Figures E.6a, E.7a and E.8a) and underrepresentation of negative sentiment (Figures E.6b, E.7b and E.8b).

- The second group of differences in network structure, between 6 of June 2021 and 11 of July 2021, can be connected to an overrepresentation of negative sentiment in the data slices starting on 27 of June 2021 and 4 of July 2021.

- In the third group, the new group we identify with this set of parameters, we find little evidence of change in sentiment during the time windows corresponding to these networks.

- The final group, that starts at the beginning of October 2021 and stretches until the end of November 2021, is connected to an overrepresentation of neutral sentiment (Figures E.6c, E.7c and E.8c).

In summary, if we look at the groups of differences in network structure, we are able to observe that the same differences correspond to the same changes in sentiment toward vaccination. This further indicates that the relationship between differences in network structure and changes to sentiment are properties of the system we are studying rather than coincidental connections from emerging from a specific partition of the dataset.

E.3 Discussion

We have made an informal sensitivity analysis to the data granularity parameters used to split our datasets. Deciding the temporal window to aggregate counts is an open problem in time series analysis and the decision is often backed by subjective arguments stemming from the domain the data is acquired from. Measures have been proposed to automatize this choice, but they often rely on seasonality or stationarity of the time series (see, for example, reference [2]), properties that the time series we showed in Figures 4.3 and 4.4 do not possess.

We showed an alternative parametrization to split the data from the vaccination forum where consecutive data slices shared a greater proportion of the data than the previous parameters. We expected that this difference would increase the difficulty to identify changes between data slices from consecutive time windows, however the difference was not significant. We had previously observed a difference in networks before and after 2015, which the new parametrization further emphasized. Perhaps because the networks prior to 2015 were considered more similar to each other due to sharing more data, we found more differences in network structure after 2015, as a greater proportion of temporally contiguous data slices lead to networks whose comparison distance was greater than the median distance.

On the other hand, we observed more stability in the sentiment change, as expected. Even though the number of data slices with significant sentiment change only decreased from 20 to 18, these data slices were harder to identify. For example, if we consider changes to neutral sentiment with the new parameters, in any given data slice there were only significantly different values in two of threads, posts or users, whereas with the previous parameters we had some data slices where all three of these presented significantly different values. In spite of this discrepancy between the two parametrizations, we find the same correlation between changes to sentiment and differences in network structure.

When deciding the parameters to split the COVID-19 dataset, we had the choice of two equally plausible combinations of parameters. We opted for the combination that would help with reasoning about results, with each data slice lasting a month. In this appendix, we whether the results of our analysis would change if we had opted for smaller data slices instead. With a smaller data granularity, we are able to observe greater variability in network structure. Whereas before, all the networks seemed very similar save for the networks corresponding to the period before COVID-19 vaccines were approve, with the new granularity we are able to observe this initial difference but

also bigger network differences after July 2021. However, even if the network distance distribution is different, the periods of time when we observe distances between consecutive time windows greater than the median distance match between the two ways to split the data.

The findings in this appendix indicate that the connection between differences in network structure and change to sentiment are a property of the datasets we studied, rather than a spurious correlation stemming from a fortunate choice of parameters to split the dataset.

Appendix F

Covasim model parameters

Parameter name	Parameter description	Value
Population size	Number of agents in the simulation	100,000
Initial infections	The number of agents initially infected	100
Number of days	Number of time steps, i.e., length of the simulation	180
Number of imports	Number of daily new infections from a source outside the population of study	Pois(5)
Testing		
LFT specificity	Probability of a negative test given that the person is not infected	99.7%[185]
Bad swab rate	Proportion of PCR and LFD tests returning negative regardless of infection status	10%
Test sensitivity	Probability of a positive test given that the person is infected	Test sensitivity curves based on [61]
PCR test delay	Number of days for the test results to be known	Pois(1.2)
Isolation		
ILI prevalence	Percentage of the population infected with ILI symptoms per day. These symptoms persist for Lognormal$(6.0, 2.0)$ days	1%
Isolation period	Number of days someone isolates after testing positive	10
Isolation factor	Percentage reduction of transmission rate parameter per layer while in isolation	Household: 0%, School and Work: 90%, Community: 90%
Contact tracing		
Contact tracing probability	Proportion of contacts traced per layer after a positive test	Household: 100%, School and Work: 45%, Community: 15%
Contact tracing time	Number of days for contacts to be traced	Household: 0, School and Work: 1, Community: 2

Table F.1: Parameters modified from Covasim defaults.

Parameter name	Parameter description	Value
	Asymptomatic testing	
Testing frequency	How often people in the regular tester cohort test	Every 2 days; twice per week; once per week
Cohort size	What percentage of the population is testing frequently on LFD tests	Varies from 0 to 100% in jumps of 10%
Anti testing group size	What percentage of the population not in the regular testing group that refuses to test without any symptoms	10%
Intermittent testing group size	What percentage of the population not in the regular testing group that tests at a quarter of the rate of the regular testers	90%
Isolation on symptom onset	Adherence to isolation when developing COVID-19 or ILI symptoms	40%
Daily testing probability if symptomatic	Probability of taking up a test per day if symptomatic with COVID-19 or ILI symptoms	40%; test is PCR in Section 5.3.1 or LFD in Section 5.3.3
Test uptake after contact tracing	Adherence to taking up a test when notified by tracing of a positive contact	40%; test is PCR in Section 5.3.1 or LFD in Section 5.3.3
	Reduction of number of contacts	
Contact reduction	Proportion of non-household contacts removed from the contact networks	Varies from 0 to 100% in jumps of 5%
Isolation on symptom onset	Adherence to isolation when developing COVID-19 or ILI symptoms	40%
Daily testing probability if symptomatic	Probability of taking up a PCR test per day if symptomatic with COVID-19 or ILI symptoms	40%
Test uptake after contact tracing	Adherence to taking up a PCR test when notified by tracing of a positive contact	40%

Table F.2: Range of values for parameters varied in the asymptomatic testing experiment.

Parameter name	**Parameter description**	**Value**
	Symptomatic testing	
Isolation on symptom onset	Adherence to isolation when developing COVID-19 or ILI symptoms	Varies from 0 to 100% in jumps of 20%
Daily testing probability if symptomatic	Probability of taking up a test per day if symptomatic with COVID-19 or ILI symptoms	Varies from 0 to 100% in jumps of 20%; test is PCR in Section 5.3.2 or LFD in Section 5.3.3
Test uptake after contact tracing	Adherence to taking up a test when notified by tracing of a positive contact	Varies from 0 to 100% in jumps of 20%; test is PCR in Section 5.3.2 or LFD in Section 5.3.3

Table F.3: Range of values for parameters varied in the symptomatic testing experiment.

Appendix G

Additional Covasim results

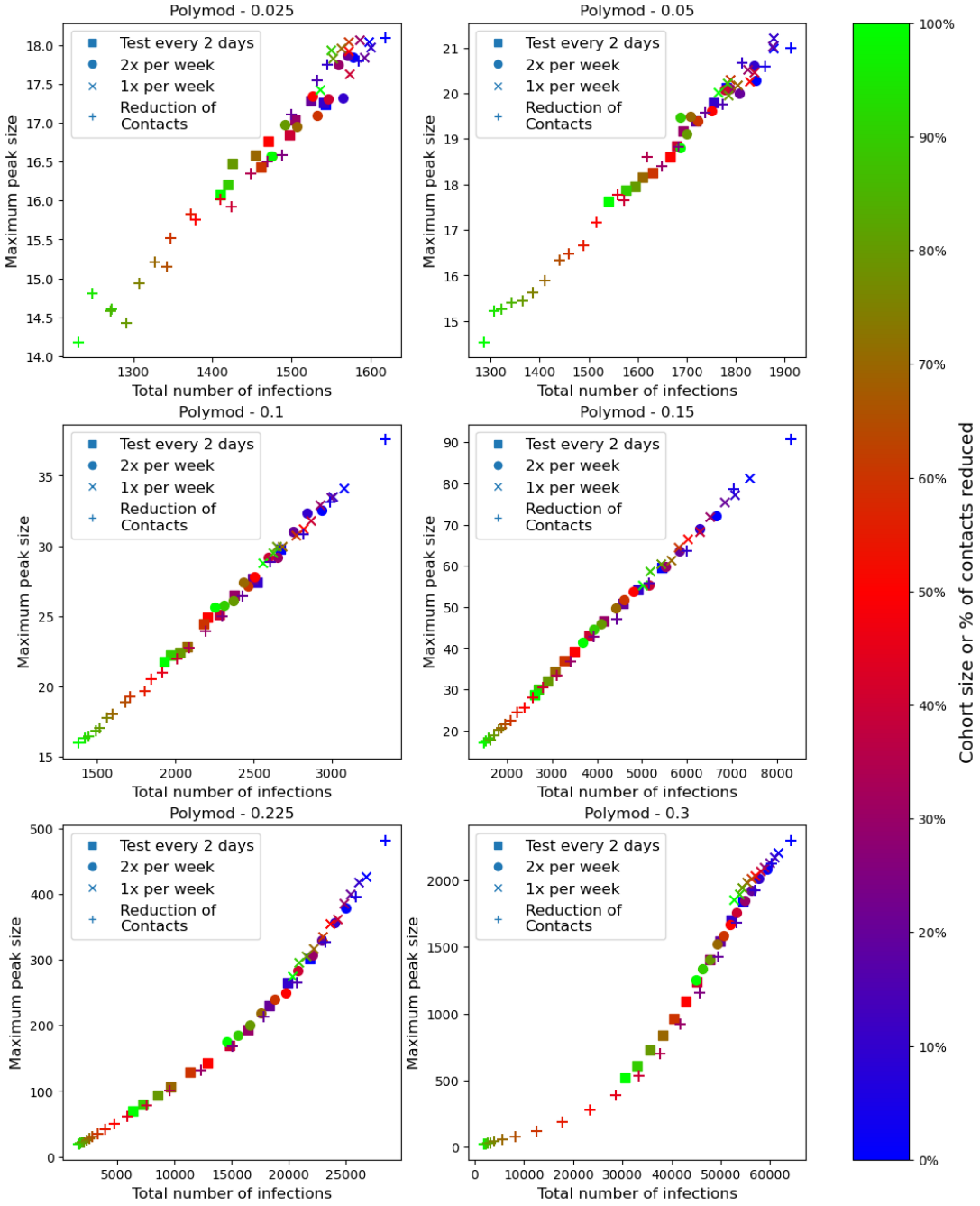


Figure G.1: Comparison of testing strategies against reduction of contacts for different growth rates with POLYMOD contact patterns. The x-axis represents the number of total infections, averaged over 100 runs. The y-axis shows the maximum peak of new daily infections, averaged over 100 runs. The colour of each marker indicates the size of the regular tester cohort for the simulations with asymptomatic testing or the percentage of contacts removed from non-household layers for the simulations with contact reduction.

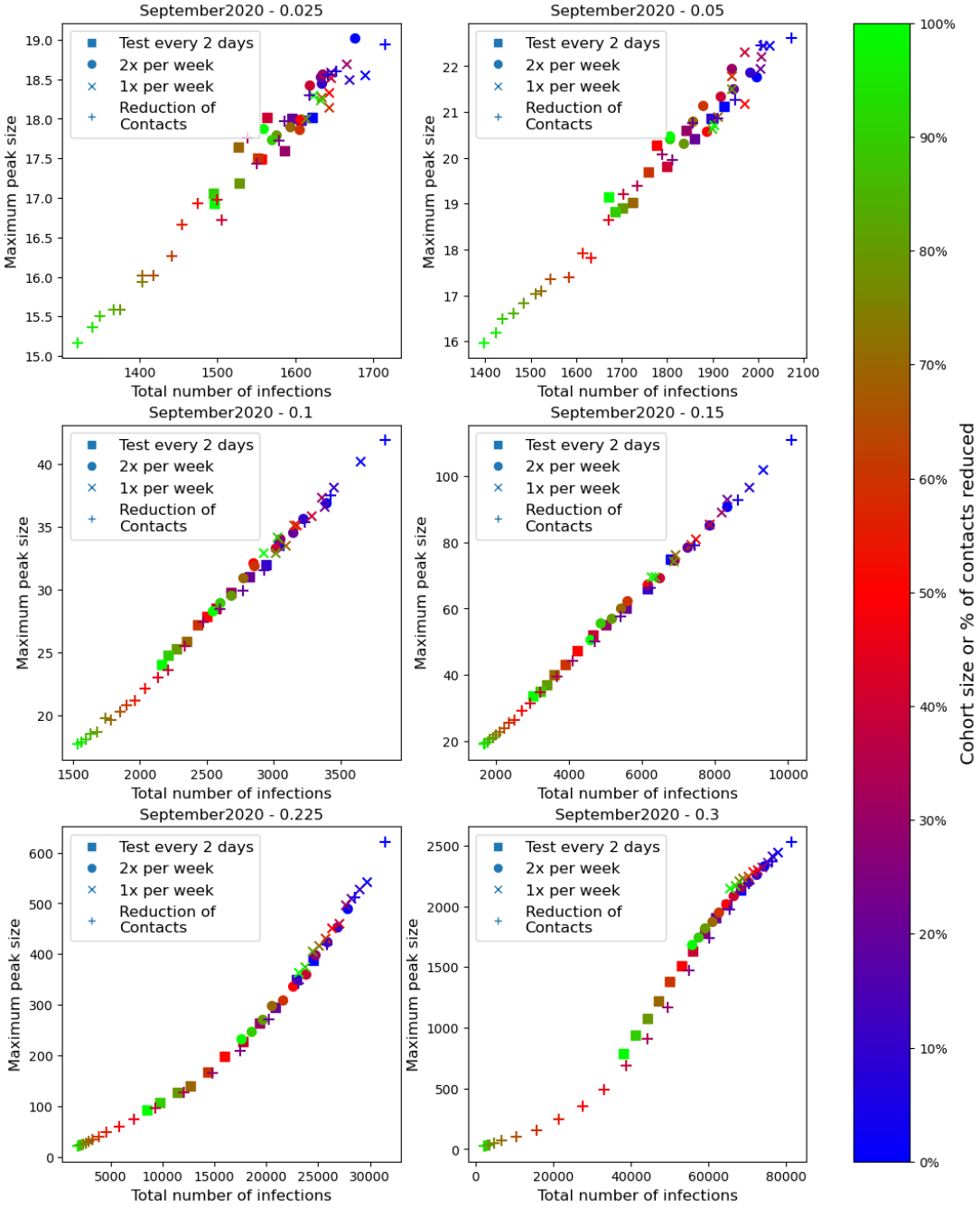
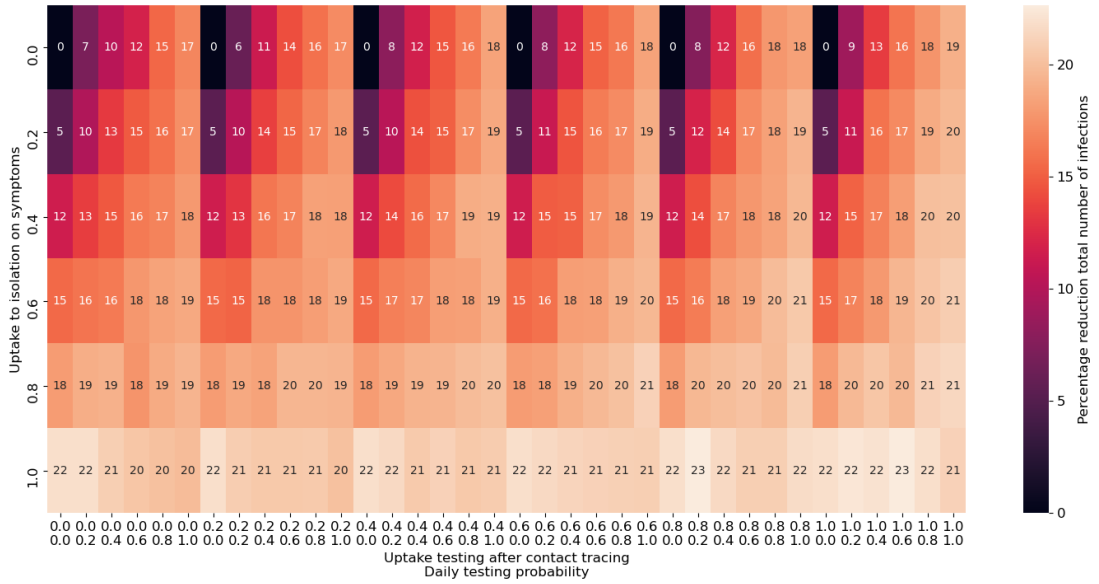
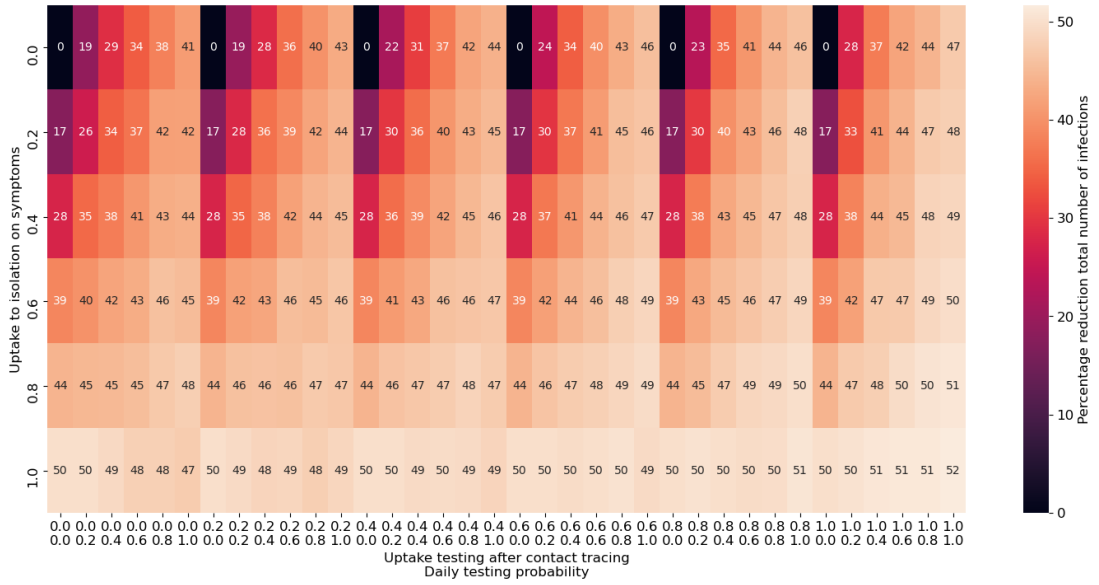


Figure G.2: Comparison of testing strategies against reduction of contacts for different growth rates with CoMix contact patterns. The x-axis represents the number of total infections, averaged over 100 runs. The y-axis shows the maximum peak of new daily infections, averaged over 100 runs. The colour of each marker indicates the size of the regular tester cohort for the simulations with asymptomatic testing or the percentage of contacts removed from non-household layers for the simulations with contact reduction.

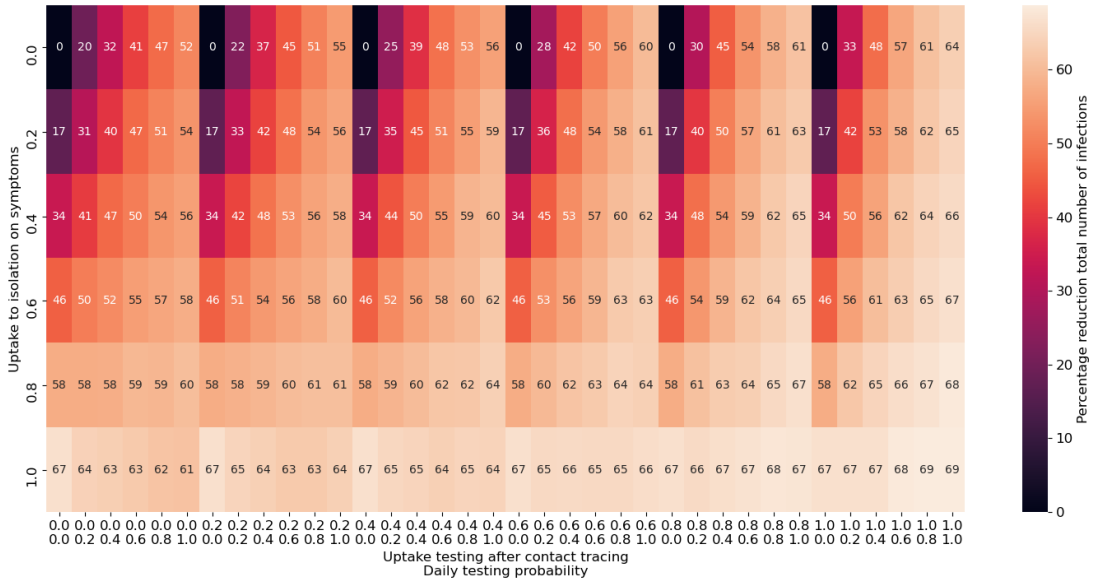


(a) Growth rate = 0.05

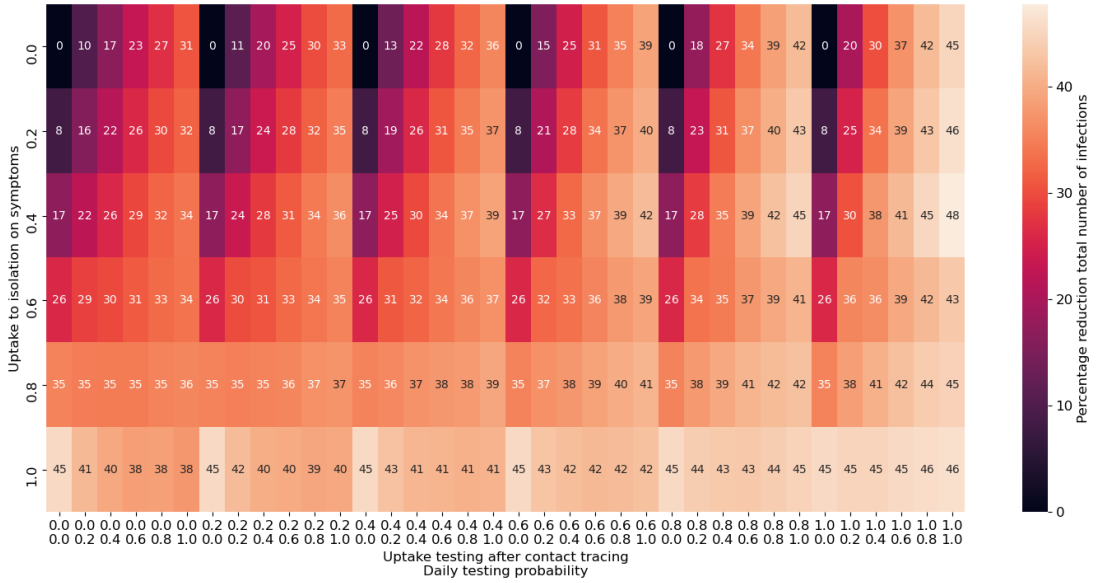


(b) Growth rate = 0.1

Figure G.3: Impact of increasing the uptake on isolation on symptoms, daily symptomatic testing and Polymerase Chain Reaction (PCR) testing after being traced as a contact of a positive case. Numbers shown are the percentage decrease in the total number of infections, compared to 0% uptake on these interventions, for 0.05 and 0.1 growth rates in POLYMOD contact pattern. The top line on the x axis represents the uptake to testing after contact tracing, the bottom line represents the daily probability to testing when symptomatic. The y axis represents uptake to isolation on COVID or ILI symptoms.



(a) Growth rate = 0.15



(b) Growth rate = 0.225

Figure G.4: Impact of increasing the uptake on isolation on symptoms, daily symptomatic testing and PCR testing after being traced as a contact of a positive case. Numbers shown are the percentage decrease in the total number of infections, compared to 0% uptake on these interventions, for 0.15 and 0.225 growth rates in POLYMOD contact pattern. The top line on the x axis represents the uptake to testing after contact tracing, the bottom line represents the daily probability to testing when symptomatic. The y axis represents uptake to isolation on COVID or ILI symptoms.

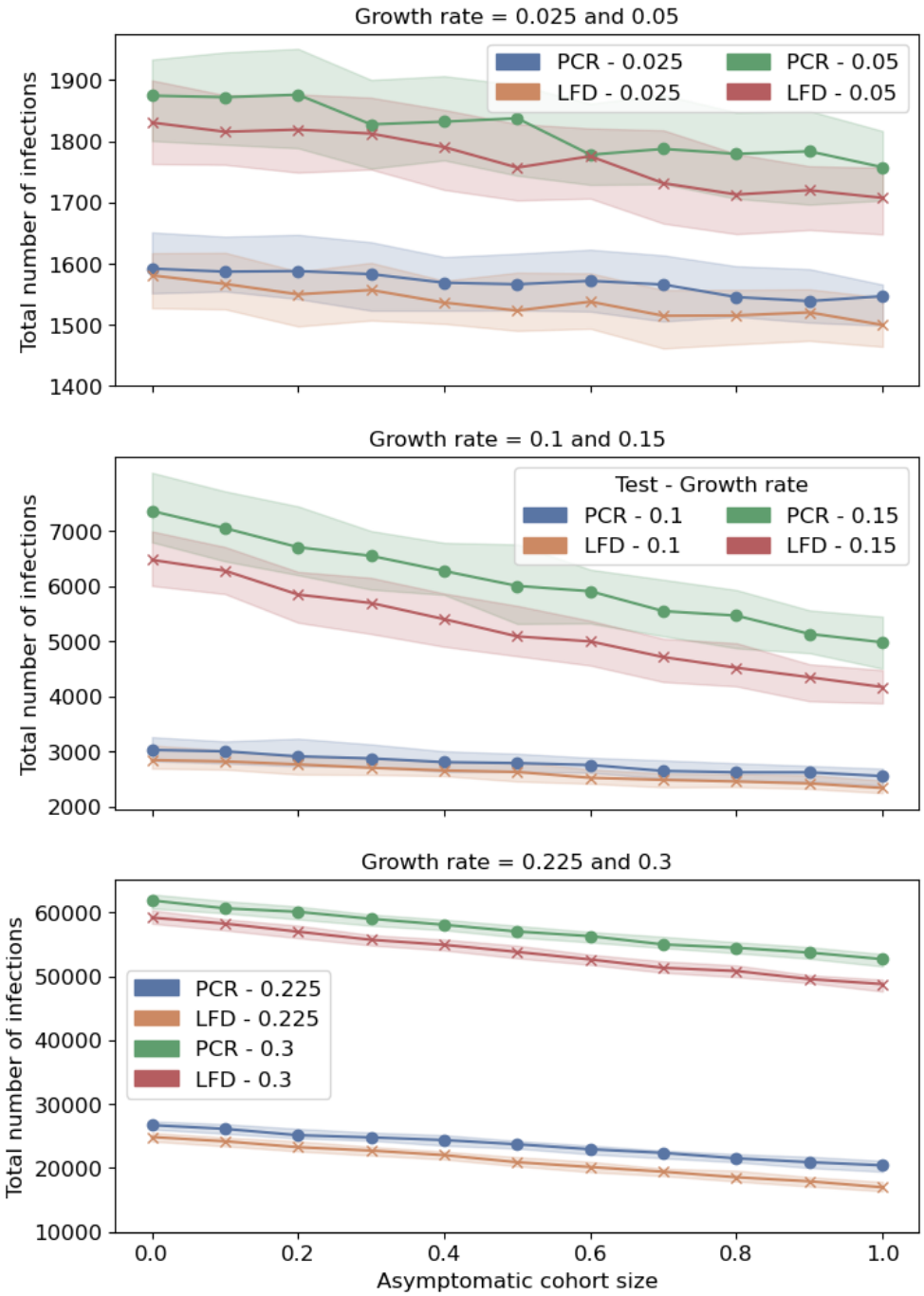


Figure G.5: Comparison of Lateral Flow Device (LFD) against PCR tests for symptomatic or contact traced individuals, for different values of asymptomatic testing uptake and different growth rates, with contact patterns obtained from POLYMOD. The y-axis shows the total number of infections. The x-axis represents the size of the regular asymptomatic tester cohort. Markers shown represent the median of 100 simulations and the shaded area represents the interquartile range. The policy used for the regular tester group is testing once weekly.

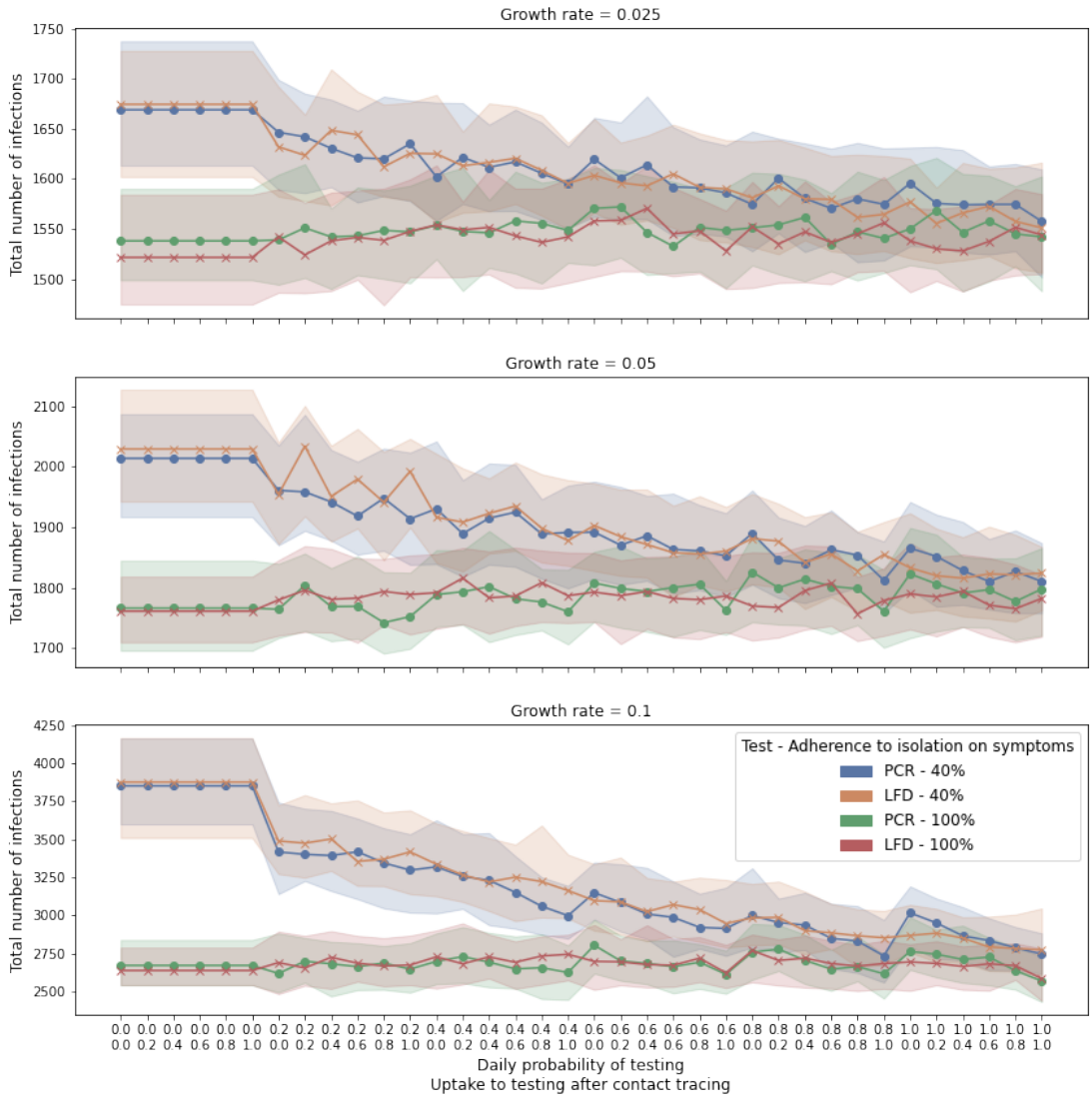


Figure G.6: Comparing LFD and PCR tests and the impact of increasing the uptake of testing after being traced as a contact of a positive case and the daily probability of testing if symptomatic, for growth rates of 0.025, 0.05 and 0.1 with contact patterns obtained from POLYMOD. The y-axis shows the total number of infections. The bottom line on the x axis represents the uptake to testing after contact tracing, the top line represents the daily probability to testing when symptomatic. Color indicates the type of test and the adherence to isolation on symptoms. Markers shown represent the median of 100 simulations and the shaded area represents the interquartile range.

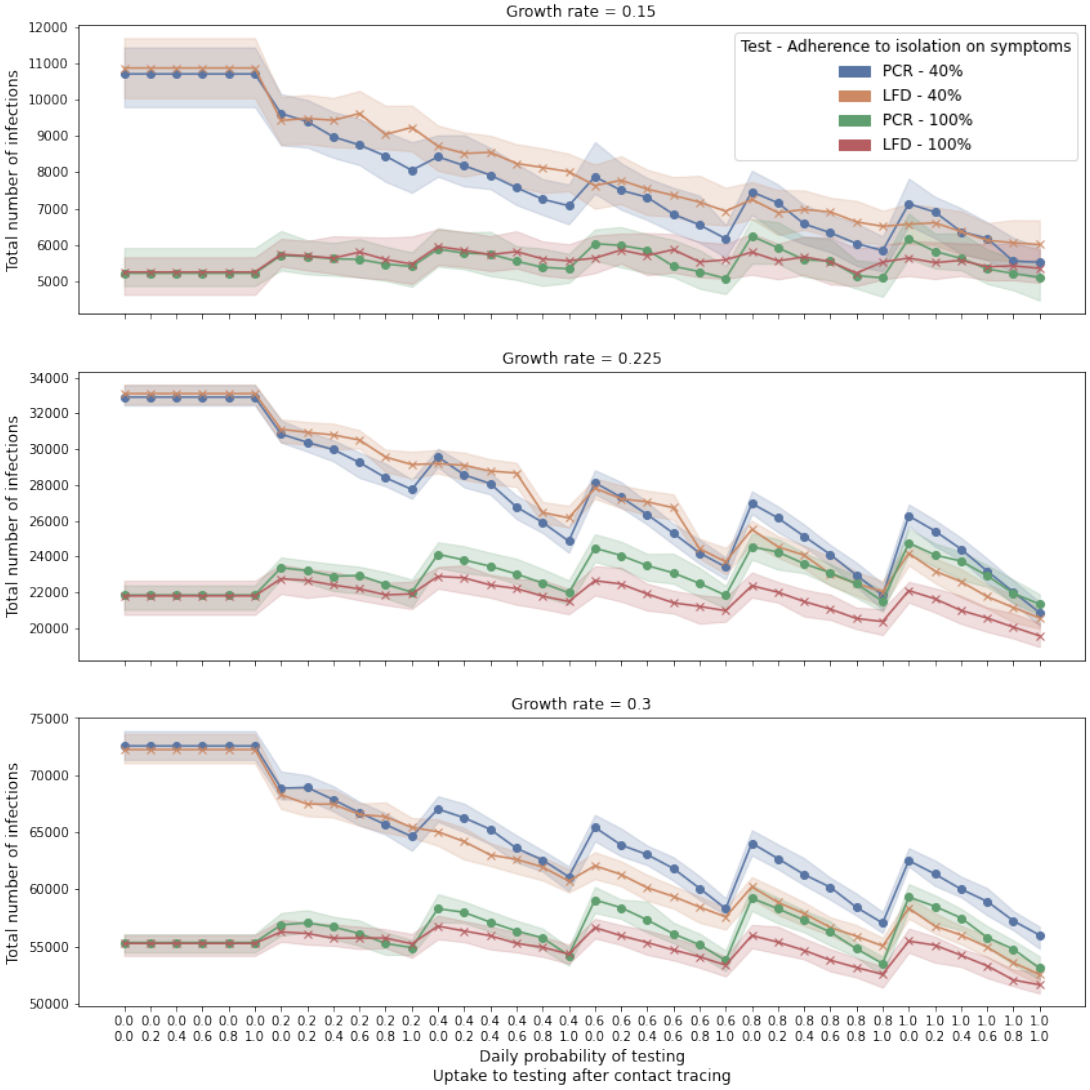


Figure G.7: Comparing LFD and PCR tests and the impact of increasing the uptake of testing after being traced as a contact of a positive case and the daily probability of testing if symptomatic, for growth rates of 0.15, 0.225 and 0.3 with contact patterns obtained from POLYMOD. The y-axis shows the total number of infections. The bottom line on the x axis represents the uptake to testing after contact tracing, the top line represents the daily probability to testing when symptomatic. Color indicates the type of test and the adherence to isolation on symptoms. Markers shown represent the median of 100 simulations and the shaded area represents the interquartile range.

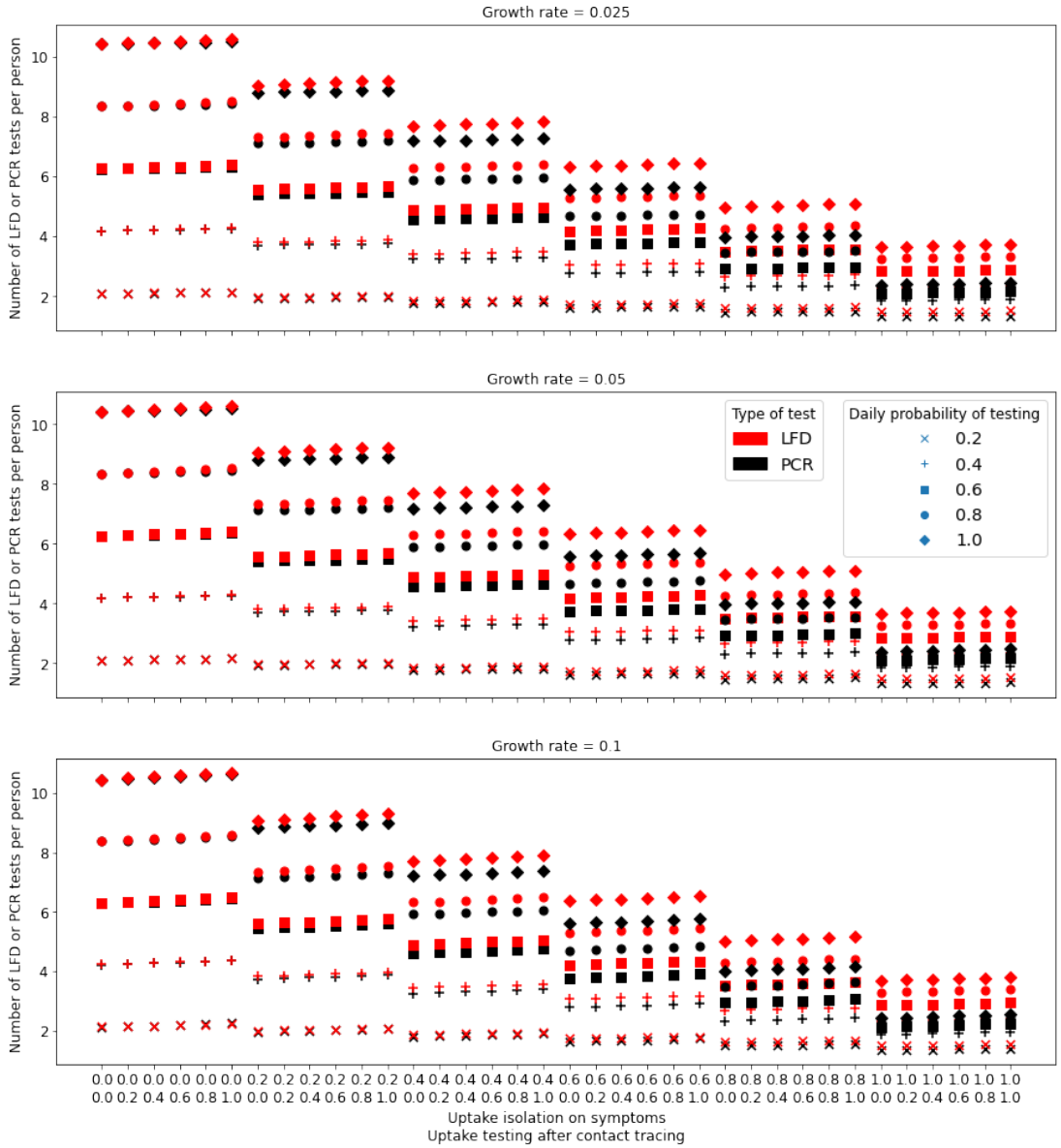


Figure G.8: Comparison of number of tests required per person when using LFD or PCR tests for symptomatic or contact traced individuals, for 0.025, 0.05 and 0.15 growth rates, with contact patterns obtained from POLYMOD. The y-axis shows the mean number of tests per person, averaged over 100 runs. The bottom line on the x axis represents the uptake to testing after contact tracing, the top line represents the adherence to isolation on symptom onset.

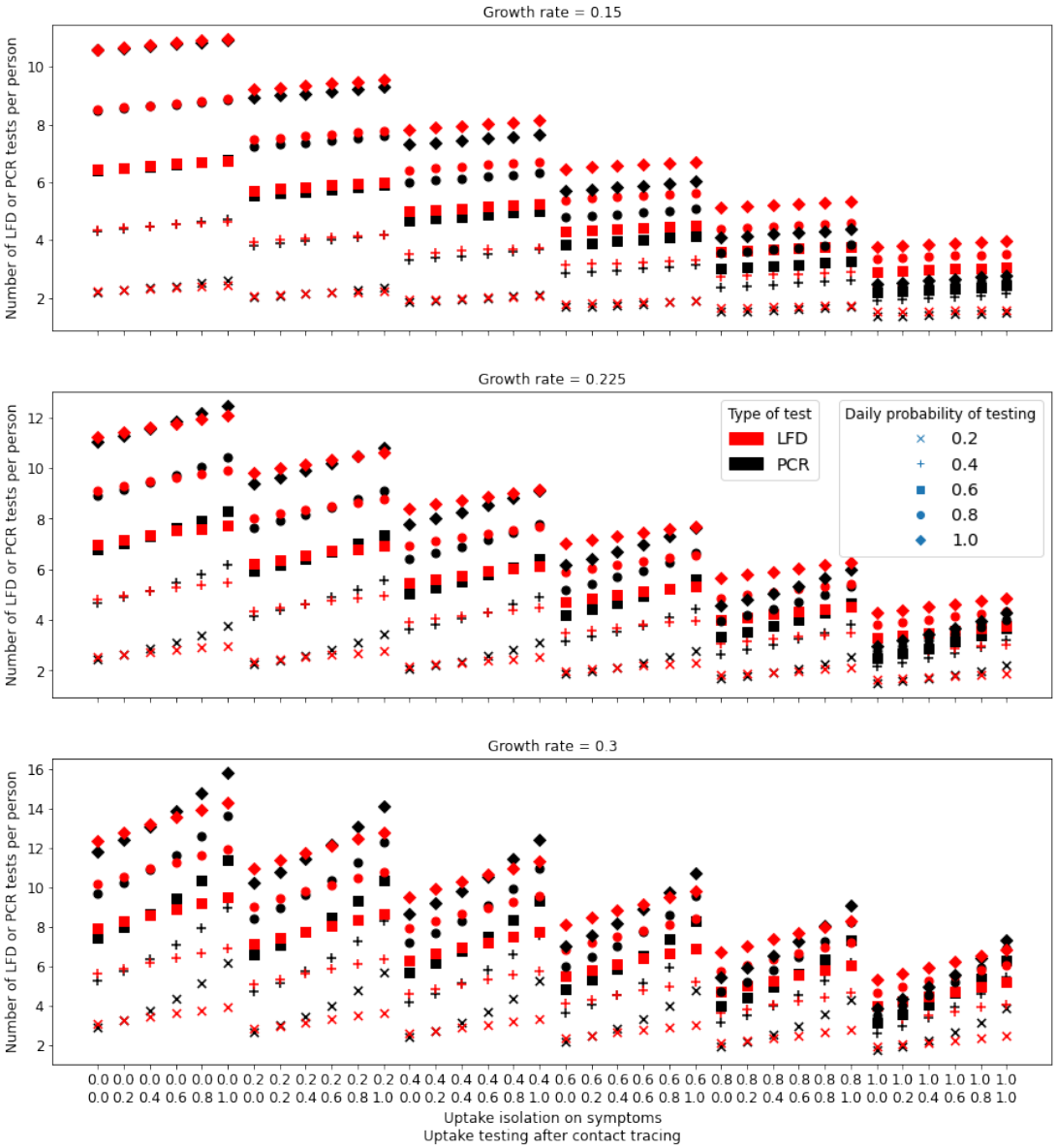
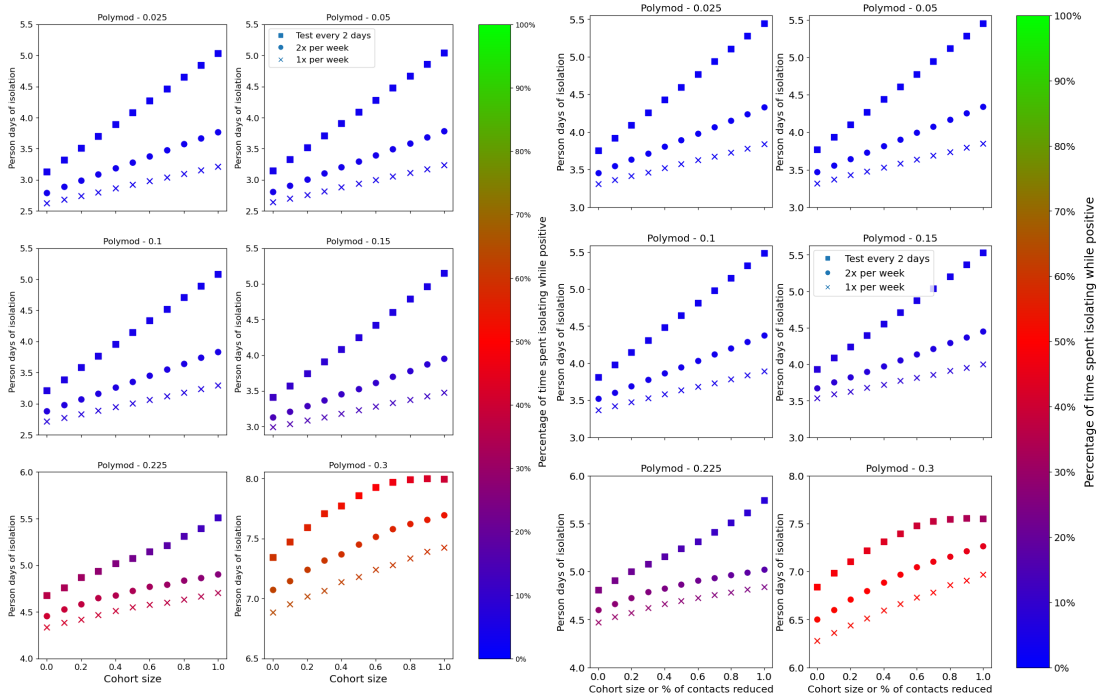


Figure G.9: Comparison of number of tests required per person when using LFD or PCR tests for symptomatic or contact traced individuals, for 0.15, 0.225 and 0.3 growth rates, with contact patterns obtained from POLYMOD. The y-axis shows the mean number of tests per person, averaged over 100 runs. The bottom line on the x axis represents the uptake to testing after contact tracing, the top line represents the adherence to isolation on symptom onset.



(a) PCR as the test for symptomatic or contact traced individuals

(b) LFD as the test for symptomatic or contact traced individuals

Figure G.10: Impact on person-days of isolation of increasing adherence to the different testing strategies, for different growth rates with contact patterns from POLYMOD, when using PCR (Figure G.10a) or LFD (Figure G.10b) as the test for symptomatic or contact traced individuals. The x-axis represents the size of the regular tester cohort for the simulations with asymptomatic testing or the percentage of contacts removed from non-household layers for the simulations with contact reduction. The y-axis shows the mean number of person days of isolation per population members over the 180 days of simulation, averaged over 100 runs. These include days in isolation waiting for test results (amongst true positives and true negatives); isolation days amongst those testing positive who were true positive; and isolation days among those testing positive who were false positives. The colour of each marker indicates the percentage of days spent in isolation while infected.

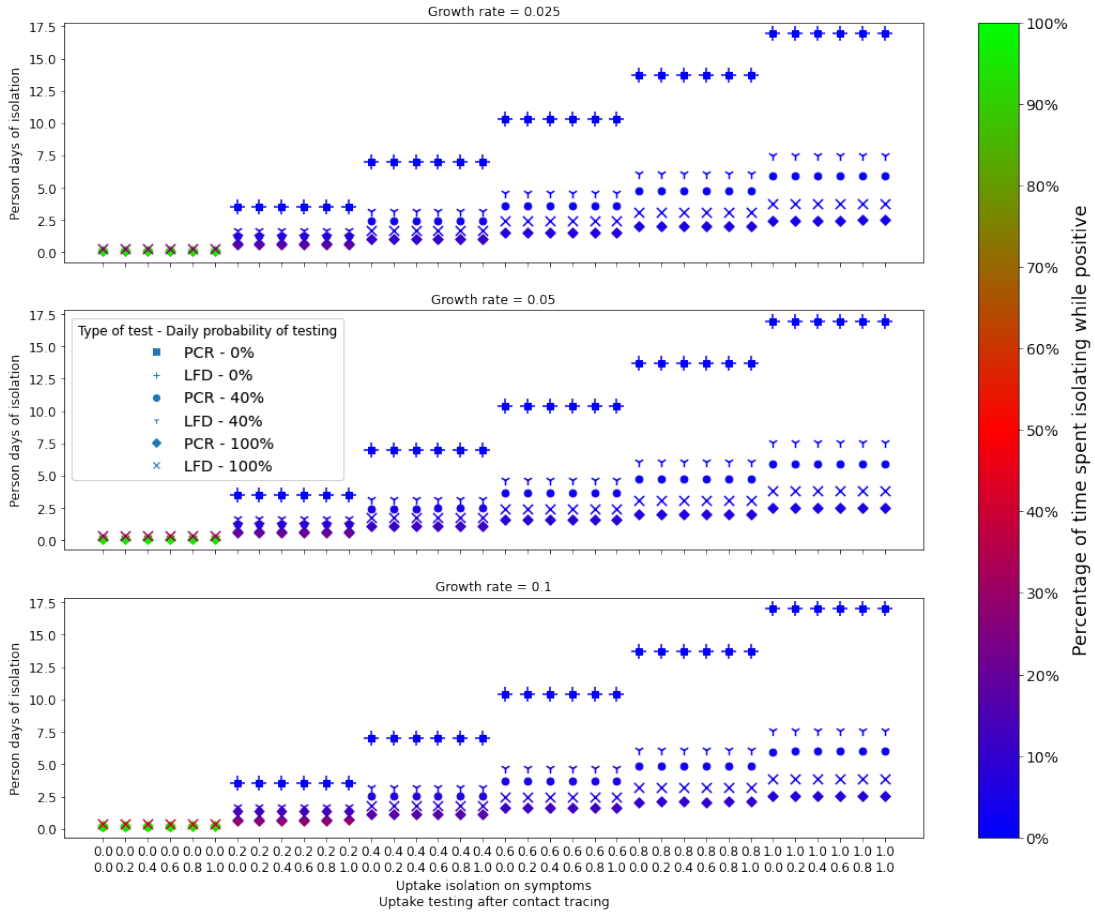


Figure G.11: Impact on person-days of isolation of increasing adherence to isolation on symptoms, daily probability of testing and adherence to testing after contact tracing, for 0.025, 0.05 and 0.1 growth rates with contact patterns from POLYMOD, comparing the outcome when testing with PCR or LFD. The bottom line on the x axis represents the uptake to testing after contact tracing, the top line represents the adherence to isolation on symptom onset. The y-axis shows the mean number of person days of isolation per population members over the 180 days of simulation, averaged over 100 runs. These include days in isolation waiting for test results (amongst true positives and true negatives); isolation days amongst those testing positive who were true positive; and isolation days among those testing positive who were false positives. The colour of each marker indicates the percentage of days spent in isolation while infected.

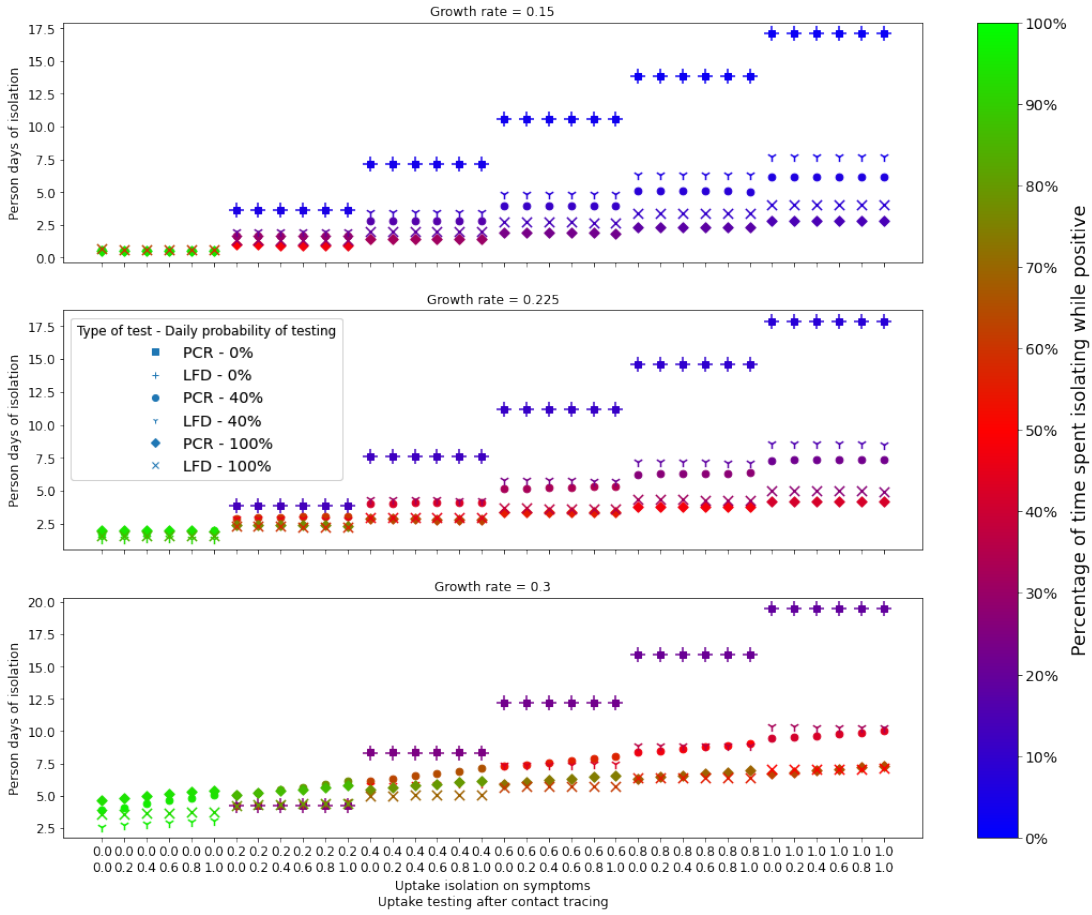


Figure G.12: Impact on person-days of isolation of increasing adherence to isolation on symptoms, daily probability of testing and adherence to testing after contact tracing, for 0.15, 0.225 and 0.3 growth rates with contact patterns from POLYMOD, comparing the outcome when testing with PCR or LFD. The bottom line on the x axis represents the uptake to testing after contact tracing, the top line represents the adherence to isolation on symptom onset. The y-axis shows the mean number of person days of isolation per population members over the 180 days of simulation, averaged over 100 runs. These include days in isolation waiting for test results (amongst true positives and true negatives); isolation days amongst those testing positive who were true positive; and isolation days among those testing positive who were false positives. The colour of each marker indicates the percentage of days spent in isolation while infected.