A NEW PRETOPOLOGICAL WAY OF IDENTIFYING SPREADERS IN PROPAGATION DIFFUSION PHENOMENA

Author(s) / Auteur(s) :

Julio LABORDE
CHArt Laboratory EA 4004, EPHE, PSL Research University, Paris, France and Insight Signals
julio.laborde@etu.ephe.psl.eu

Marc BUI
CHArt Laboratory EA 4004, EPHE, PSL Research University and University Paris 8, Paris, France
marc.bui@ephe.psl.eu

Abstract / Résumé :

In a world that's increasingly connected, many crises are related to propagation phenomena where we need to either repress the spreading (e.g. epidemics, computer viruses, fake news...) or try to accelerate it (e.g. the diffusion of a new anti-virus patch). A good understanding of such phenomena involves a knowledge of both the structure of the whole system and the specifics of the transmission process. The standard way to deal with the former has been through a characterization of the structure by the use of networks, where nodes are the components of the system where the propagation occurs, and links exist between them if there's a possibility of transmission from one component to the other. This allows to identify the super-spreaders (i.e. components that diffuse in a disproportionally large amount) as nodes with certain particular network properties. Here we propose the use of pretopology as a framework to characterize the structure of a system, as well as a new pretopological metric for the identification of super-spreaders. Since the metric can easily be transformed into an equivalent network metric, it is easy to compare its performance with some of the classical network indices of node importance. The relevance of the metric is tested by the use of some standard agent-based models of epidemics and opinion dynamics. Finally, a pretopological model of opinion diffusion is also proposed and studied.

Keywords / Mots-clés :

pretopology, networks, centrality, diffusion

INTRODUCTION

Problems related to diffusion are everyday more present in our lives (epidemics, missinformation spreading, computer viruses, etc...). To confront these problems the strategy has been to identify the individuals[1] that are more central in the system in order for them to stop or accelerate the diffusion process. But that begs the question of what is to be central in a complex system. This question is strictly conditioned to the way the system structure is characterized, and the traditional way to deal with this has been by the use of networks (Wasserman, S. & Faust, K. (1994), Amblard, F. & De□uant, G. (2004), Newman, M. E. J. (2004), Borgatti, Stephen P. (2005)). In this work, we reaffirm the pertinence of pretopology as an alternative framework for descriving a structure.

Once the framework has been set, we still have many possible metrics to quantify how central an individual is and is far from obvious to decide which of those metrics are more relevant.

One way to justify the comparative advantage of a metric over the others has been the use of agent-based models. This paradigm, where one models the individual behavior of agents and studies the global properties that emerge from their interaction, has proven extremely flexible when it comes to model, and extremely rich in its results (Hegselmann, R. (2002), Hegselmann, R, & Krause, U. (2005), Hu et al.(2015)). In these models agents are usually a simplified representation of individuals of a population, with properties and methods representing the features and actions that characterize those individuals.

[1] We will mostly talk about individuals or agents, having the context of a social system in mind, but the whole framework could be applied to any other system with multiple components where a process of diffusion might occur.

Since the models are usually quite simple, they allow to make multiple simulations, with many sets of parameters, without using to many computational resources. The idea is not that much to find the correct paramaters for the model, but to better understand the relations between those parameters, the dynamics of the system and the emergent properties.

We will make use here of a set of agent-based models to test for the effectiveness in identifying central agents of both network and pretopological metrics. The "real" centrality of an agent will be measured by the amount of influence he/she can exersice over the rest of a population during a simulation.

CENTRALITY METRICS

We will start by briefly remembering the concepts that are used by both network theory and pretopology to identify central nodes, and then we will proceed to introduce our new metric.

Frameworks

Graph Theory

Network theory takes its power from graph theory (Bollobás, B. (1998)). A graph G(V, E) is a mathematical object composed of a set V of nodes, and a set E of edges, i.e. pairs of nodes that are connected. When using a graph to describe a system, we associate the components with nodes, and edges with connections or relations between the components.

Graph theory proposes a series of metrics that allow to quantitavilly differenciate nodes according to their structural position, or to differenciate among different graphs.

Some of these metrics are (see Fig.1):

- **Degree of a node:** the number of neighbors a node has in the graph, that is, the number of nodes that are connected to the node. In a Social network, a node with a high degree represents an individual that knows a lot of people, and for that same reason he's likely to be an important person.
- The **betweenness centrality** of a node is computed by taking the shortest paths between every couple of nodes in the graph, and calculating the fraction of them that includes that node; if the fraction is high it would mean that when two nodes want to communicate in an optimal way (through their shortest path), then there's a good chance that they will pass through this node. A node with this characteristics could have capital importance if we wanted to hinder the communication inside the network.
- The **clustering coefficient** of the graph tries to measure how likely is for two neighbors of a node to be neighbors among them. This metric is usually important because it allows to uncover a phenomenon that we find very frequently on empirical networks; indeed, it's a lot more likely for two friends of someone to be acquainted among them, than for two randomly selected people.

Graphs can also be expressed in therms of matrices, and this other representation also allows to get important insights into the structure of the population through techniques of matrices decomposition or spectral analyses. This is the case, for example of the eigenvector centrality or the pagerank index.

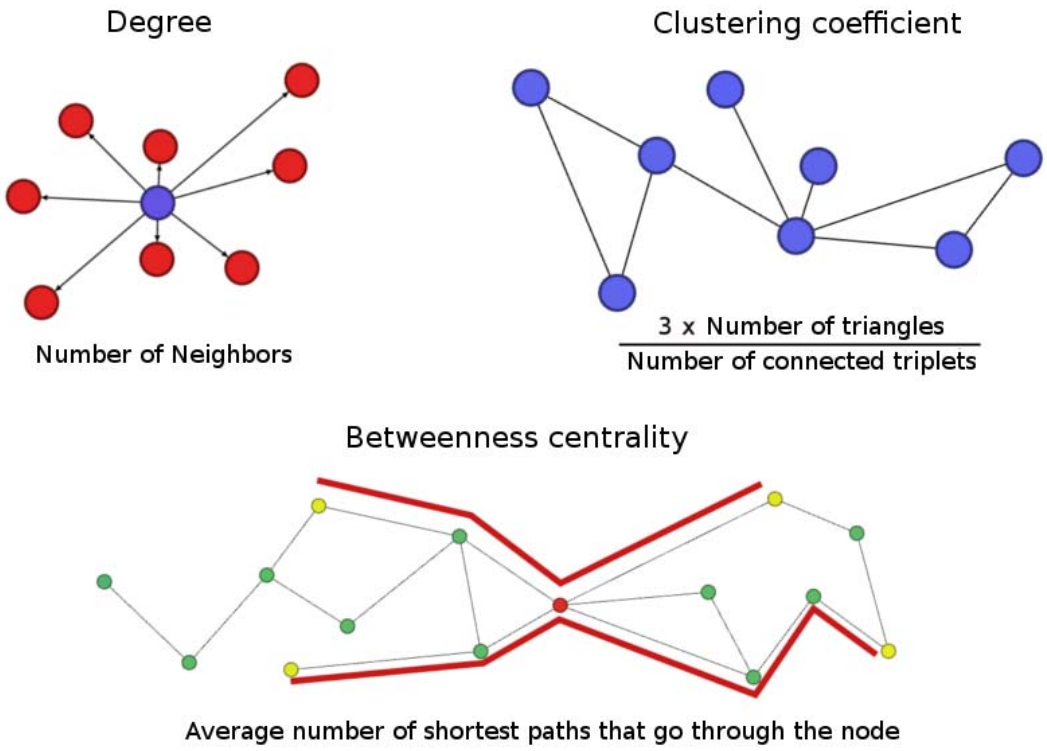


Figure 1. Graph metrics

Pretopology

Another way of characterizing a system is by the use of pretopology (Belmandt, Z. (2011))

A pretopology on a set U is defined by a function $a:\square(U)\to\square(U)$, such that:

- $a(\square)=\square$ *(Preservation of Nullary Union)*
- $\forall A\in\square(U), A\subseteq a(A)$ *(Extensivity)*

We shall call that function a pseudoclosure function on U.

If we also ask that:

- $\forall A,B\in\square(U), A\subseteq B\Rightarrow a(A)\subseteq a(B)$ *(Isotonic)*

we get what is called a *V-type* pretopology.

Finally, if:

- $\forall A,B \in \wp(U), a(A \cup B) = a(A) \cup a(B)$ *(Additive)*

we get a V_D*-type* pretopology.

It's interesting to realize that all these definitions are different degrees of relaxation of Kuratowski's axioms for a topology, where it was also demanded that $\forall A \in \wp(U), a(A) = a(a(A))$ (Idempotence).

Pretopology has been used in the context of social systems (Levorato, V. (2014), Bui, Q.V. *et al.* (2018)), where it has been shown that it's a well suited theory to model phenomenons where the relations existing among groups of agents are not just the sum of the relations among the individual agents. This is quite a normal situation when dealing with social phenomena.

Since the number of possible subsets of a set is 2^N , where N is the number of elements it contains, then even for a small set it becomes infeasible to completely list the image for each of the subsets under the pseudoclosure function. In practice, instead of explicitely presenting the pseudoclosure of every subset, people have usually characterized a pretopology by a set of rules to construct the pseudoclosure of a given subset.

This algorithmic description has done that preopology has been used to model the evolution of a system in time, where each application of the pseudoclosure function models one time-step of the system. In this paper, on the other hand, we are interested in a static interpretation of pretopology, where the pseudoclosure function gives a structure to the set of agents that's permanent over time.

Teambuilder index

When using pretopology to model a system or a diffusion inside of a system, we are normally interested in groups that have a large pseudoclosure. We propose here to concentrate on the individuals that on average will increase the most the pseudoclosure of a group.

We start by defining the function *group_teambuilder()* as:

$$group\_teambuilder(x,A) = |a(A \cup \{x\})| - |a(A)|$$

Where $|A|$ is the cardinality of the set A .

In other words, we calculate the pseudoclosure of a set not containing an element x , and then we calculate the pseudoclosure of the union of that set with x , if the difference between the two is large, then the element x made a big contribution. By doing this over all possible sets, we are able to identify the members that on average contribute the most to enlarging a set.

That is to say we are looking for the elements x that have the largest *teambuilder* index, which is defined as follows:

$$teambuilder(x) = \sum_{A \subseteq U | x \notin A} group_{teambuilder}(x,A) = \sum_{A \subseteq U | x \notin A} |A \cup \{x\}| - |A|$$

As mentioned before, a naive approach to this calculation would involve a sum of as many as 2^{N-1} elements for a set U with N elements, which becomes quickly intractable. While this might be the only way to calculate the index for any possible pretopology, we will concentrate here in a particular case.

It has been customary (Belmandt, Z. (2011)) to derive pretopologies from one or more networks defined over a same set U, by deciding the membership of an element x to the pseudoclosure of a set $A \subseteq U$ according to some rules applied to the edges of those networks between x and the elements of A.

We propose here a fast and practical way to calculate the *teambuilder* index on the simplest case where the pseudoclosure is defined over a set U such that there's one unidirectional, unweighted network $G(U,E)$ defined on U. For this case the pseudoclosure is defined in the following way:

$$x \in a(A) \Leftrightarrow (x \in A) \vee (\exists y \in U | (x,y) \in E)$$

Put differently, an element x belongs to the pseudoclosure of a set A, if the set contains x or contains at least one graph neighbor of x.

Under these circumstances the *teambuilder* index is given by the following formula:

$$teambuilder(x) = \sum_{y \in neigh(x)} 2^{N - neigh(y) - 1}$$

To see why this is so, we need to change from asking *"for each set* A *such that* $x \notin A$*, how many elements would add* x *to the pseudoclosure?"*, to ask *"for each element* z*, how many sets* A *such that* $x \notin A \wedge z \notin a(A)$ *will have z in their pseudoclosure after joining x (i.e.* $z \in a(A \cup \{x\})$*).* The answer to this second question is simply *"all those sets that don't already have* z *in their pseudoclosure"*, and these are just the sets that don't contain z or any of the neighbors of z.

In practice we won't need to calculate those powers of 2, since we will only be interested in comparing the difference of *teambuilder index* between elements, and this can be done by just stocking the different exponents of 2 that we are adding. So the teambuilder index can be calculated for all x in *O(E)*.

Although it is a fundamentally pretopolical measure, the teambuilder index is associated to each node, and for the case of a pretopology built according to a network it can easily been translated into a network metric.

In Figure 2 we can appreciate how the team builder is a sort of compromise between degree and betweennes centrality. Indeed, the highest teambuilder is not necessarily the person with more contacts, since for a very clustered network, the addition of that node to a group wouldn't add many new neighbors if the group was already connected to most of its contactcs.

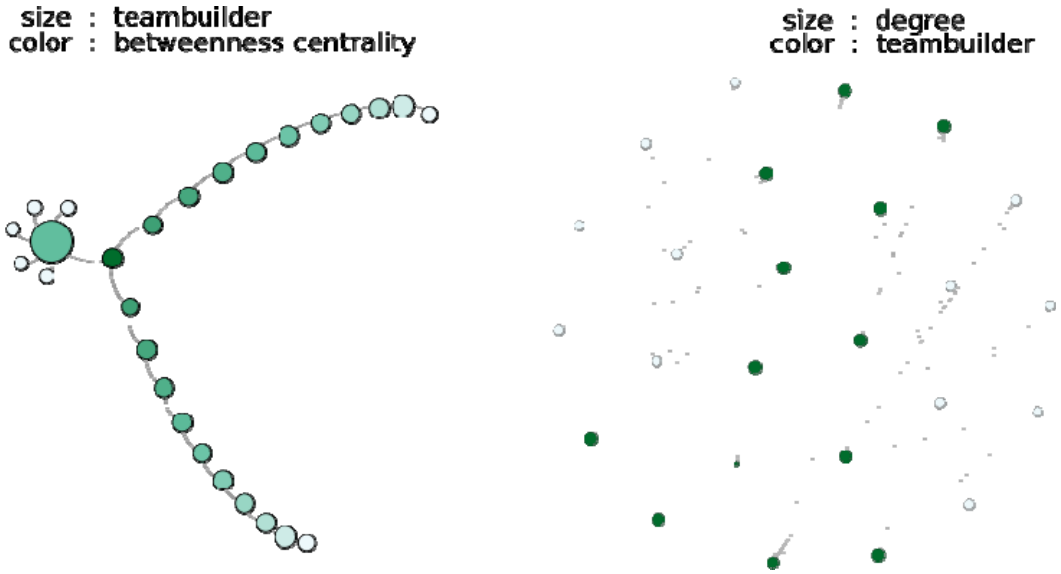


Figure 2. Difference between the teambuilder index and other graph metrics. On the left, difference with betweenness centrality, on the right, difference with degree.

SIMULATIONS

We will now introduce the different models that will be used to test how well the teambuilder index captures the idea of an agent being central in a process of diffusion. Although these models are presented as opinion dynamic models, they are so general that have been used as models of epidemics, segregation and other diffusion phenomena (Ben-Zion, Y. et al. (2010), Volz, E., & Meyers L.A. (2007), Schelling, T. C. (1971))

Models

Network models

The methodology consists in building a population, create a network on that population, select a target group that will have a different opinion than the rest, and execute an opinion dynamic model.

The following three models of opinion dynamic will be used:

- A Cascade model
- A Threshold model
- A Utility-based model (Hu et al.(2015))

The creation of the network is done according to the same procedure used in (Hu et al.(2015)). This procedure allows to control for the degree of clustering of the network through a parameter called *homophily_level,* which at the same time captures the idea that people are usually connected to people that are similar to them.

For each value of the parameter *homophily_level* different networks will be created (parameter *number_networks*); for each network, all selection strategies will be performed; if the selection strategy for the targets is not deterministic, we will select a target group and execute the opinion dynamic models multiple times (parameter *number_selections*). The results are stocked and averaged over all networks that used the same *homophily_level*, and all selection tries for those networks.

It's important to notice that only in the Utility-based model the homophily-level will be really modeling the concept of homophily, since the algorithm for network creation will use the motivation of each agent to connect them (c.f. bellow), but none of the other models use that variable. So for those other two models the homophily_level will only control the structural clustering.

The parameters and the pseudocode for the whole process are the following:

Parameters:

- population_size = 1000
- average_neighbors = 10
- number_networks = 15
- number_selections = 15
- homophily_levels = [0, 0.2, 0.4, 0.6, 0.8, 1.0]
- targets: 50

```
function methodology():
        create_agents()
        for hl in homophily_levels
                for i=1 to number_networks:
                        build_network(hl)
                        for selection_strategy in selection_strategies:
                                if selection_strategy is random:
                                        for j in number_selections:
                                                select_targets()
                                                cascade_diffusion()
                                                threshold_diffusion()
                                                utility_diffusion()
                                else :
                                        select_targets()
                                        cascade_diffusion()
                                        threshold_diffusion()
                                        utility_diffusion()


function create_population()
        for i = 1 to population_size:
                agent_i.motivation ← random(-1, 1)     // motivation is used by the utility model
                agent_i.conformity ← random(0, 1)      // conformity is used by the utility model
                agent_i.threshold ← random(0, 1)       // threshold is used by the threshold model


function build_network(homohily_level):
        for i = 1 to number_of_edges:
                n1 ← random_node()
                j ← random(0, 1)
                if j < q:
                        n2 ← select_closest(n1)    // The node with the closest motivation
                else:
                        n2 ← random_node()
                add_edge(n1, n2)
                edges[n1, n2].influence ← random(0, 0.2)  // Each edge has an influence
```

Opinion dynamic models:

Each simulation starts with the targets having opinion 1 and the rest of the population having opinion 0.

```
function cascade_diffusion():
    for t in steps:
        for n in population:
            if n.opinion_t == 1                 // n's opinion changed
             and n.opinion_t-1 == 0:            // previous step
                for n2 in neigh(n):
                    if n2.opinion == 0:
                        x ← random(0, 1)
                        if x < edge[n, n2].threshold :
                            n2.opinion_t+1 = 1
```

```
function threshold_diffusion():
    for t in steps:
        for n in population:
            if sum(neighbors(n).opinion == 1)/|neighbors(n)|        // fraction of neighbors
             > threshold:                                           //   with opinion 1
                n.opinion = 1
            else:
                n.opinion = 0
```

```
function utility_diffusion():
    for t in steps:
        for n in population:
            if n.conformity(1 - 2*sum(neighbors(n).opinion == 1)/|neighbors(n)|)
             + (1-n.conformity)*n.motivation  >  0:        // the utility of 1 is bigger
                n.opinion = 1                              //    than the utility of 0
            else:
                n.opinion = 0
```

We finally list the different selection strategies, with an explanation when it's pertinent:

- Random selection
- Largest degree
- Smallest degree
- Largest conformity
- Smallest conformity

- Largest motivation
- Smallest motivation
- Largest teambuilder index
- Smallest teambuilder index
- Group_teambuilder heuristic: we test an heuristic for selecting a large pseudoclosure by selecting a random node x_1, and then selecting the node x_2 that maximizes group_teambuilder(x, $\{x_1\}$). Subsequently we add the node x_3 that maximizes group_teambuilder(x, $\{x_1, x_2\}$). The process continues until we have selected the number of targets necessary.
- Largest eigenvector centrality
- Largest pagerank centrality
- Random pseudoclosure: we take many random sets of targets, we measure the size of the their pseudoclosures, and we select the one with the largest one. The number of random sets selected was fixed to 500 in our simulations; the motivation being to spend around three times the amount of time spent with the group_teambuilder heuristic.
- Largest betweenness centrality
- Largest closeness centrality

Mixed model

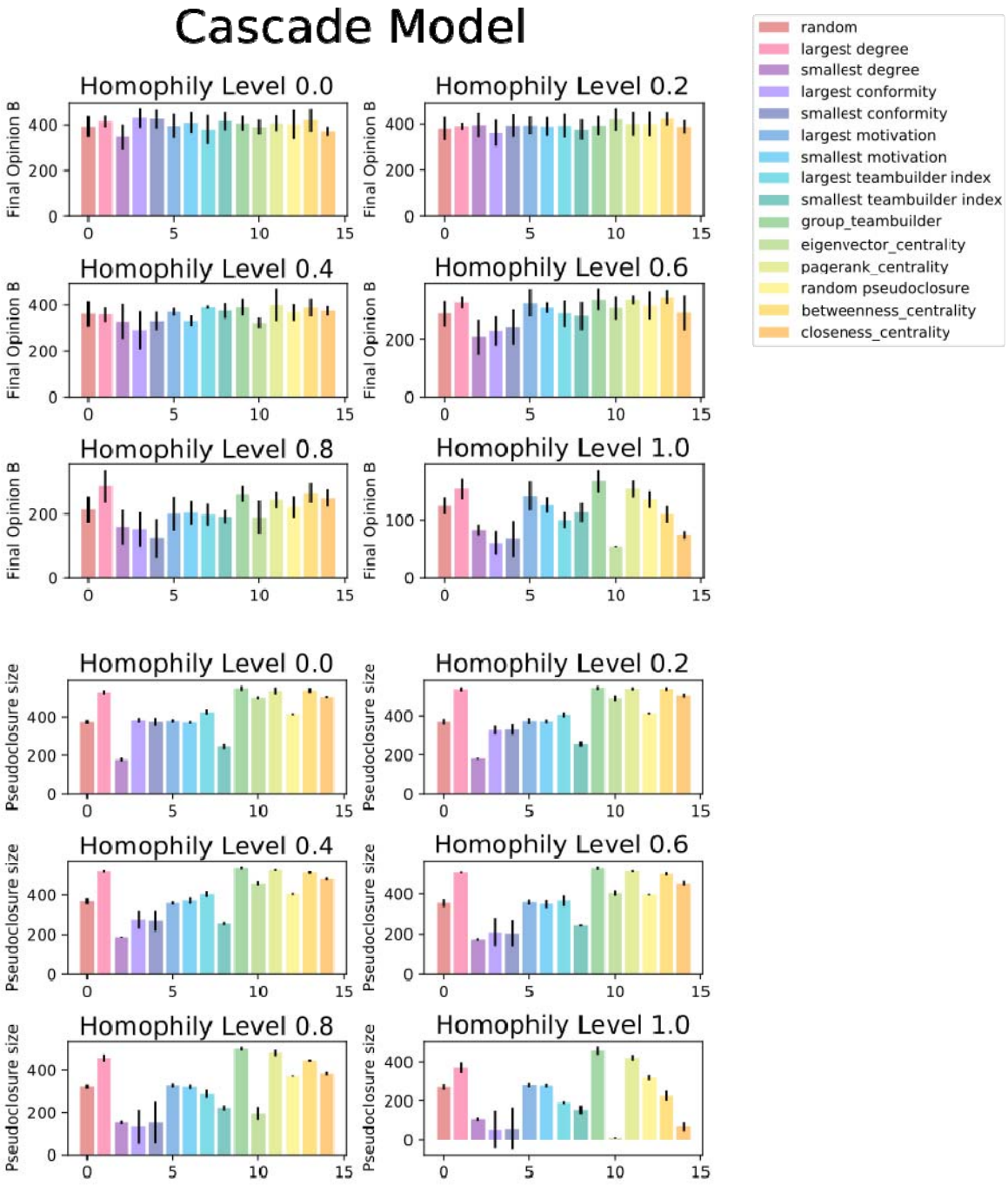
A second type of model is presented where after creating the social network from the previous models, we also a create an enemy network. This is a random network with average connectivity of 5.

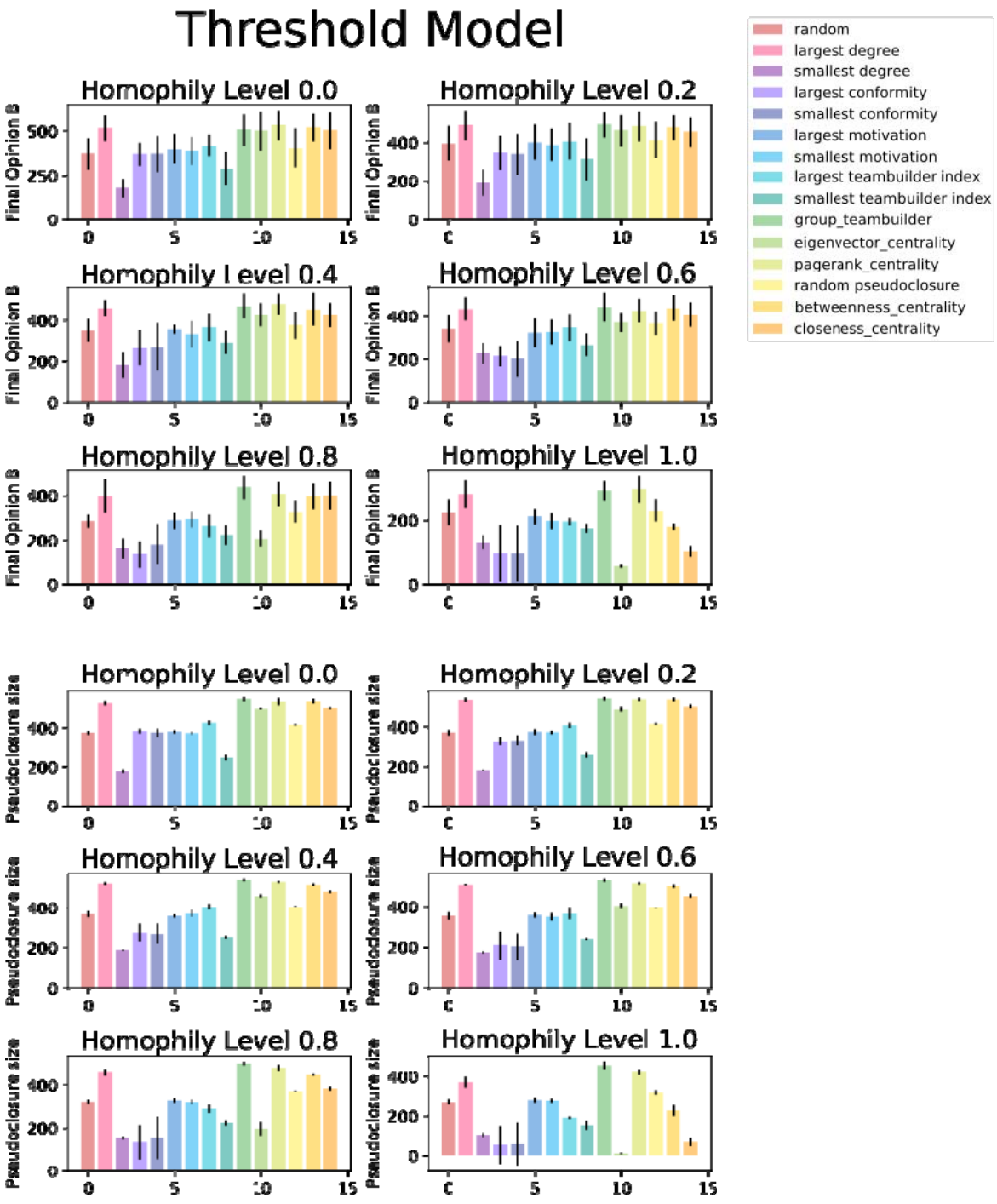
The opinion dynamic is the same as the threshold model, only in this case every enemy with opinion 1, substractcs a friend with opinion 1. We call this model the *mixed model.*

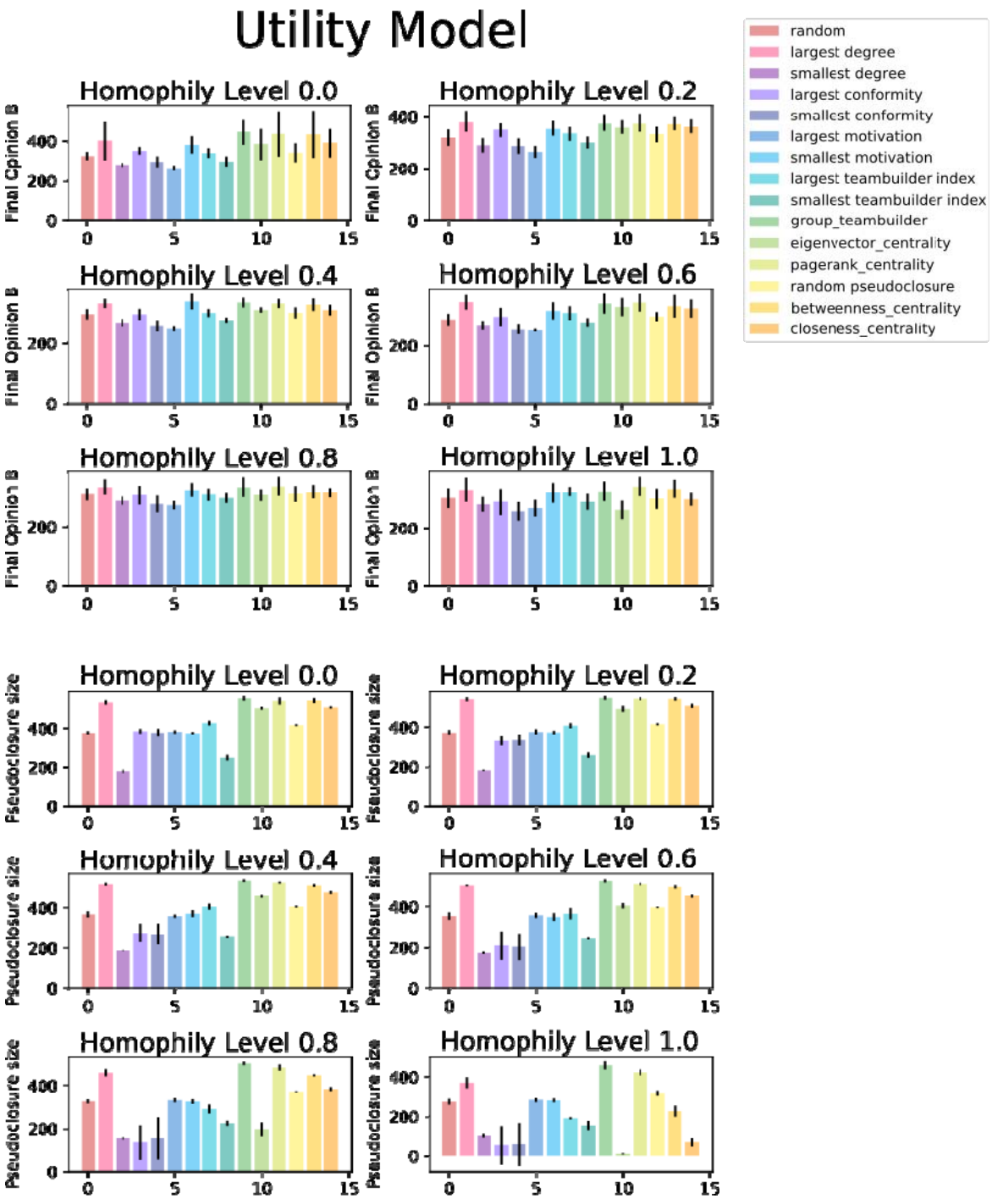
In order to better study this second model we propose a second type of pseudoclosure that will define a new pretopological space over the set of agents; here an agent x will belong to the pseudoclosure of a set A if the number of friends (i.e. connections in the first network) inside the set are more than twice the number of enemies inside the group. The idea behind this is to find a set that is not only well connected, but also connected in a friendly way, so its impact on the opinion of its neighbors should be possitive.

In order to differentiate this second pseudoclosure from the first one, we shall call it mm_pseudoclure, and the corresponding target strategies mm_group_teambuilder and mm_random_pseudoclosure.

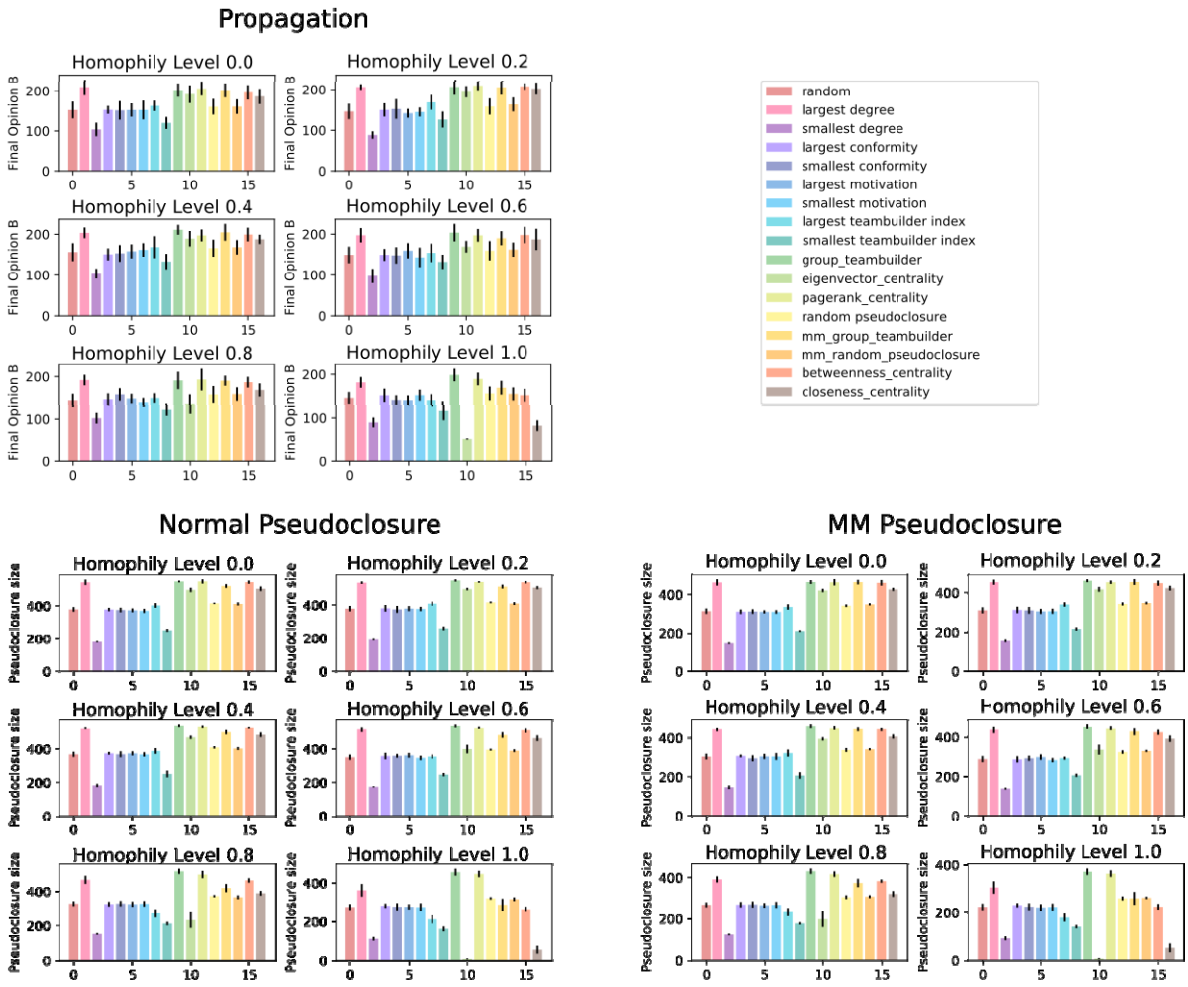
Results


Cascade Model
Homophily Level 0.0
Homophily Level 0.2
Homophily Level 0.4
Homophily Level 0.6
Homophily Level 0.8
Homophily Level 1.0
Final Opinion B
Pseudoclosure size
random
largest degree
smallest degree
largest conformity
smallest conformity
largest motivation
smallest motivation
largest teambuilder index
smallest teambuilder index
group_teambuilder
eigenvector_centrality
pagerank_centrality
random pseudoclosure
betweenness_centrality
closeness_centrality


Threshold Model
random
largest degree
smallest degree
largest conformity
smallest conformity
largest motivation
smallest motivation
largest teambuilder index
smallest teambuilder index
group_teambuilder
eigenvector_centrality
pagerank_centrality
random pseudoclosure
betweenness_centrality
closeness_centrality
Homophily Level 0.0
Homophily Level 0.2
Homophily Level 0.4
Homophily Level 0.6
Homophily Level 0.8
Homophily Level 1.0
Final Opinion B
Pseudoclosure size


Utility Model
Homophily Level 0.0
Homophily Level 0.2
Homophily Level 0.4
Homophily Level 0.6
Homophily Level 0.8
Homophily Level 1.0
Final Opinion B
Pseudoclosure size
random
largest degree
smallest degree
largest conformity
smallest conformity
largest motivation
smallest motivation
largest teambuilder index
smallest teambuilder index
group_teambuilder
eigenvector_centrality
pagerank_centrality
random pseudoclosure
betweenness_centrality
closeness_centrality

Results Mixed Model



Analysis

Individual network models

The first thing we note is that the more clustering we have, the more different are the metrics among them, with most of them performing in a very similar manner when the network is completely random, but degree centrality, pagerank and group_teambuilder standing out as the homophily level increases, and group_teambuilder working systematically better for the cascade and threshold model.

When interpreting this result in the context of a social system and a marketing strategy, for example, it becomes interesting to realize that the degree of an agent in the world (e.g. how much "followers" he/she has) usually correlates to a bigger price or effort to convince him/her to adopt a new strategy. The pretological approach on the other, does not necessarily take agents that are individually the most powerful, but those that as a group are influential, so the individuals in the group may be cheaper or easier to persuade.

The teambuilder index on the other hand, while not irrelevant to the amount of spreading (as can be seen by the difference between the targets with the largest and smallest teambuilder indices) , did not perform particularly well as a strategy for selecting the targets.

Another thing that we can notice is how correlated are the size of the pseudoclosure with the final propagation of the opinion, proving that under the right circumstances it should suffice to study the structure of the population to get a good idea of their dynamics.

A final interesting result is that the group_teambuilder heuristic works consistently better at finding big pseudoclosures than random_pseudoclosure, although the number of random samples had been chosen so it takes around three times more to select the targets randomly than with group_teambuilder. The interest of finding good heuristics for the problem of finding large pseudoclosures becomes obvious when we realize that no polynomial time algorithm may exist unless P=NP. Indeed, we can see that by thinking that if we had a polynomial time algorithm for finding the biggest pseudoclosure, we could apply it to the sets of a single element, then to those of two elements, and follow until the biggest pseudoclosure would include the whole space; this would give a polynomial time algorithm for the *minimum dominating set problem,* a problem well known to be NP hard.

For the mixed model

Here the difference of propagation can be seen from the smallest amounts of homophily level.

The size of the different pretopologies is also correlated to the final propagation in the mixed model, proving that our intuition in defining the mm_pseudoclosure was also correct, but contrary to our beliefs neither the mm_group_teambuilder nor the mm_random_pseudoclosure performed any better than regular group_teambuilder or pagerank in identifying groups with large mm_pseudoclosure. The performance being quite similar for random networks, but getting increasingly worse as the homophily level augments.

It's worth noting though that the mm_group_teambuilder heuristic was once again much more successful than mm_random_pseudoclosure in identifying groups with large mm_pseudoclosure.

CONCLUSION

In this work we have restated the validity of pretopology as an alternative framework for the structuring population (or any other system with many parts), we have presented the teambuilder index, a new pretopological metric, and we have shown that for a particular case it can be calculated exactly in only O(E) time.

Our most interesting results come from the heuristic based on the teambuilder notion that we developed subsequently, which not only showed to be a better selection strategy for spreaders than all the network-based ones, but perhaps more importantly, seems to be a much better way to uncover sets with a large pseudoclosure than a simple random strategy, and this using a lot less resources. This last result is particularly important if we want to impose pretopology as a pertinent choice for structuring systems in a world where those systems and the data describing them are becoming ever larger.

RÉFÉRENCES

Amblard, F. & De□uant, G. (2004). “The Role of Network Topology on Extremism Propagation with the Relative Agreement Opinion Dynamics,”.

Banisch, S, & Olbrich, E. (2017). “Opinion Polarization by Learning from Social Feedback,”.

Bollobás, B. (1998). *Modern Graph Theory*. Graduate Texts in Mathematics 184. New York: Springer.

Bansal, S., Read J., Pourbohloul B. & Meyers L.A. (2010). “The Dynamic Nature of Contact Networks in Infectious Disease Epidemiology.” *Journal of Biological Dynamics* 4 (5): 478–89.

Belmandt, Z. (2011) *Basics of Pretopology*. Hermann.

Ben-Zion, Y., Yahel Cohen, and Nadav M. Shnerb. (2010). “Modeling Epidemics Dynamics on Heterogenous Networks.” *Journal of Theoretical Biology* 264 (2): 197–204.

Borgatti, Stephen P. (2005). “Centrality and Network Flow.” *Social Networks* 27 (1): 55–71.

Borgatti, S.P. & Everett M.G. (1992). “Notions of Position in Social Network Analysis.” *Sociological Methodology* 22: 1.

Bui, Q.V., Ben Amor, S. & Bui, M. (2018) Stochastic Pretopology as a Tool for Topological Analysis of Complex Systems

Cohen, R, Havlin S. & ben-Avraham D. (2003). “Efficient Immunization Strategies for Computer Networks and Populations.” *Physical Review Letters* 91 (24).

Deffuant, G., Neau D., Amblard, F. & Weisbuch, G. (2000). “Mixing Beliefs among Interacting Agents.” *Advances in Complex Systems* 03 (01n04): 87–98.

Degroot, Morris H. (1974). “Reaching a Consensus.” *Journal of the American Statistical Association* 69 (345): 118–21.

Douven, I., and A. Riegler. (2010). “Extending the Hegselmann-Krause Model I.” *Logic Journal of IGPL* 18 (2): 323–35.

Granovetter, M. (1978). “Threshold models of collective behavior.” Am. J. Sociol. **83**(6), 1420–1443

Hegselmann, R. (2002) “Opinion Dynamics and Bounded Confidence Models, Analysis, and Simulation”.

Hegselmann, R, & Krause, U. (2005). “Opinion Dynamics Driven by Various Ways of Averaging.” *Computational Economics* 25 (4): 381–405.

Hu, Hai-hua, Jun Lin, and Wen-tian Cui. (2015). “Intervention Strategies and the Diffusion of Collective Behavior.” *Journal of Artificial Societies and Social Simulation* 18 (3).

Kiesling, E., Gunther, M., Stummer C. & Wakolbinger L.M. (2009). “An Agent-Based Simulation Model for the Market Diffusion of a Second Generation Biofuel.”

Levorato, V. (2014). “Group Measures and Modeling for Social Networks.” *Journal of Complex Systems* 2014: 1–10. https://doi.org/10.1155/2014/354385.

Levorato, V, & Bui M. (2010) “Modeling the Complex Dynamics of Distributed Communities of the Web with Pretopology,”.

Moore, T., Finley P., Brodsky, N., Brown, T., Apelberg, B., Ambrose, B. & Glass R. (2015). “Modeling Education and Advertising with Opinion Dynamics.” *Journal of Artificial Societies and Social Simulation* 18 (2).

Newman, M. E. J. (2004). “Detecting Community Structure in Networks.” *The European Physical Journal B - Condensed Matter* 38 (2): 321–30.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology* 1: 143-186.

Volz, E., & Meyers L.A. (2007). “Susceptible-Infected-Recovered Epidemics in Dynamic Contact Networks.” *Proceedings of the Royal Society B: Biological Sciences* 274 (1628): 2925–34.

Wasserman, S. & Faust, K. (1994). *Social Network Analysis: Methods and Applications*, Cambridge University Press

Watts, D.J. & Dodds, P.S. (2007). “Influentials, Networks, and Public Opinion Formation.” *Journal of Consumer Research* 34 (4): 441–58.