Structural Aspects of User Roles in Information Cascades

Io Taxidou, Peter M. Fischer
University of Freiburg, Germany
{taxidou,peter.fischer}@cs.uni-freiburg.de

ABSTRACT

Social media plays an important role for the exchange and dissemination of information among its users. In turn, these users shape social media by their interactions, online behaviour and status. These aspects differ massively from user to user, which has an impact on the outcome of information diffusion. However, there has been observed patterns of online user behaviour which lead to distinct *user roles*. While there has been a lot of research on user roles and information diffusion in isolation, their combination has not been researched much. In this paper we study their correlation, in particular whether particular user roles occur in specific structural positions in information cascades. By testing several hypotheses, we could confirm that there is indeed correlation of these two aspects. However, some user roles demonstrate diverse behaviour with regard to their activity patterns and need further investigation.

1. INTRODUCTION

Information diffusion researches the processes of how a piece of information propagates on social media. Such analysis considers **structural** aspects, i.e., information propagating from user to user over social graphs and *social* aspects - real users acting in the virtual space of social media. Information diffusion is modeled with *information cascades*: information cascades are graphs that reveal how information propagates from user to user, often with the assumption of an underlying social graph. Information diffusion is generated by online users who get influenced and propagate content. Their online behaviour varies which is demonstrated by their interactions and status in social media. However, particular types of user behaviour i.e. *user roles* [22, 2] have been observed in the literature[22, 2, 20], that can characterize a large share of the online population.

While research on information diffusion [8, 15] and user role identification [22, 2] in social media have each received considerable attention, the correlation of these two aspects has not been investigated much. In this paper, we research the correlation of 1) structural aspects of information diffusion and 2) user roles that are derived from online human behaviour. In particular we seek to identify *structural positions* of user roles in information cascade graphs. For example: Are celebrities mostly at the root of cascade graphs? Are spammers at the leaves because they do not trigger further reactions? By confirming (or rejecting) such hypotheses we can shed light on the mechanisms of information diffusion. By specifying the structural positions of user roles, we can make better predictions for the outcome of information diffusion: for example a particular user role that is observed more often at the leaves might signify the end of information diffusion.

The remainder of the paper is structured as follows: In Section 2 we discuss related work and in Section 3 we describe our dataset. Section 4 provides the methodology and results for reconstructing information cascades, while in Section 5 we identify prominent user roles. Section 6 investigates the correlation of information cascades and user roles by computing structural positions on information cascades. Finally, 7 concludes the paper.

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3053906

2. RELATED WORK

Information Diffusion in social media has been a board field of research. In this respect, models of information diffusion are developed like [9, 18, 7] and information cascades have been researched in mamy contexts [5, 15, 25, 6, 14] We provide some examples of analysis over information cascades. In [24] authors investigate the size, shape and decay factors of cascades during the Irianian "Green Revoution" in 2009 while in [13] shape and temporal metrics of retweet cascades were evaluated. The authors in [10] investigate human interactions on a emergency event constructing the corresponding cascades. In [12] the impact of location, time and distance is examined with regard to information adoption, and the list continues to grow.

For identifying prominent user roles we discern two categories: 1) supervised methods like in [22, 1] where a framework is used and user roles are adapted to this framework according to the defined features; 2) unsupervised methods like in [3, 20] where datasets drive the cluster creation (number of clusters not known a priori) and results need to be interpreted accordingly.

In more detail, the work in [22] develops a model based upon the Twitter message exchange to identify key players in conversations. This model categorizes Twitter users into specific roles based on their dynamic communication behavior. The work in [1] applies a semantic model and rules combined with statistical analysis in order to compute behaviour in online forum communities. This model categorises behaviour of forum community members over time, and researcher how different behaviours correlate with community growth in these forums. Analysis of user intentions in Twitter was implemented by [11], were the intention of each post was determined manually. The user intentions discovered which also categorise users include: Daily Chatters, Conversations, Sharing Information and Reporting news. For unsupervised methods,

authors in [3] cluster users in Twitter according to their activity. The number of clusters and the quality of clusters and not known in advance. In the same lines, the work of [20] clusters forum users, according to their posting behaviour and following semantic rules.

While information diffusion and user roles has individually been studied in the past, much less is known about the connection between these two. The closest to our work is [23] which studies the interplay between users' social roles and their influence on information diffusion. The model proposed integrates social roles and diffusion modeling into a unified framework. Such a model can be used to predict whether an individual user will repost a specific message in the micro level; while at the macro-level, the model can predict the scale and the duration of a diffusion process. However, our goal is different since we are trying to identify structural positions of user roles in information cascades. Complementary the work of [16] identifies communities of users on top of information cascades: in our case we decouple features to detect user roles and information cascades since we aim to identify the correlation of both. Authors in [17] correlate diffusion processes with the evolution of the underlying social graph. This problem has been adapted to a probabilistic generative model [4] that allows the understanding and reproduction of such processes.

3. DATASET

The dataset we are using was recorded on the 2012 summer Olympics in London using terms like "olympics", "london2012". It contained 13.6 M messages, 2.27 M distinct users and 1.1 M retweet cascades.

In order to obtain reliable results for computing the structural metrics, we filtered out cascades with size lower than five messages. We ended up with 4.618 cascades which is the dataset we are using to compute the metrics presented in Section 4. For identifying prominent user roles we considered 2,27 M users who contributed at least two messages in during the 2012 Olympics.

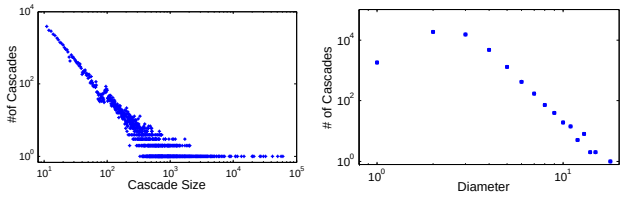
4. INFORMATION CASCADE RECONSTRUCTION

In this section, we present the methodology to reconstruct information cascades. This allows us to compute structural metrics for the cascade users that show structural positions (in Section 6) We focus on retweet cascades, but these methods can be applied to other diffusion processes (e.g. hashtags or replies) and different social media.

When users are retweeting, Twitter provides the initial source of a message (root), but not the intermediate forwarders that exposed the information to them. In other words, the intermediate diffusion paths are not provided by Twitter. Under the hypothesis that information flows through social connections (users are exposed to information from their followers), we leverage the social graph to search for possible influencers and unravel the intermediate diffusion paths. We use our algorithm from [21] that reconstructs retweet cascades, given some social graph. This algorithm allows multiple influencers in case more than one of the user's followers are retweeting the same message. As a result, retweet cascades are DAGs, with a single root. Note here that in other means of propagation e.g. hashtags we might observe multiple roots.

Figure 1 presents the distribution of cascade sizes and diameters.

As shown in figure 1a, cascades have a skewed distribution of size with the large majority yielding a few reactions. The largest cascade has around 62K of messages while around 5K cascades have more than 100.



(a) Cascade Size Distributions (b) Diameter Distribution

Figure 1: Cascade Properties [21]

Figure 1b shows that cascades tend to be deep, with a mean value of diameter 4. Diameter values up to 18 are observed, indicating that information is being propagated to large audiences beyond the root's followers. This has an impact on cascade shapes, which results in complex structures as well as star structures. Having a diversity of cascade shapes serves well our purposes of computing several structural metrics on them.

5. USER ROLE IDENTIFICATION

Next, we identity prominent user roles according to their online behaviour from the dataset in 3. We do not rely on any predefined model of influence or on any pre-knowledge of user groups as in [22]. For such analysis we need to 1) select features that characterize users and their activity and 2) a clustering algorithm that groups users with similar features together. We considered features that reveal:

- status: number of followers, number of friends, number of times being mentioned, is verified, has url
- activity and engagement: number of tweets, number of retweets, number of replies, number of mentions, number of off-topic messages - messages that do not refer to the crawled dataset according to keywords)
- ability to trigger further reactions: retweet and reply reaction rates (fraction of messages that receive at least one retweet and reply)

Note that, no information diffusion aware features were used to identify user roles. As a result, there is no beforehand correlation of user roles and structure in cascades, as in [16].

After extracting features for every user, we need to select a clustering algorithm that will reveal which distinct groups of users exist in the data. We also need to also define the number of clusters, since this information is needed by most of the clustering algorithms. We tested K-Means and Expectation Maximization (EM) as clustering algorithms; EM assigns a probability distribution to each user which indicates the probability of belonging to each of the clusters. This is very useful in case users fall between more than one cluster or their behaviour deviates over time. In order to assign each user to one cluster, we get the maximum probability for each user to identify the most "fitting" cluster. Both methods require the number of clusters k to be provided in advance.

Since we have no a priori information about the number and quality of clusters, we have to define an objective function that shows the best clustering approach and number of clusters for our dataset. Our goal is to a) maximize the cohesion of data items within each cluster and b) maximize the separation of clusters, so that we end up with well-defined clusters. In practice, the similarity of data items within each cluster, and the dissimilarity of data items among different clusters have to be maximized. The similarity

and dissimilarity can be computed by any distance metric like the Euclidean distance. We used the Silhouette coefficient [19] which accounts for these two parameters. It takes values from $[-1, 1]$ and the higher its value, the better the clustering is.

We tested several number of clusters (according to literature (e.g. [20]) in the range [3, 20] for K-means and EM. The optimal number of clusters for both methods was nine which was identified by testing the Silhouette Coefficient on the aforementioned range.

Expectation maximization yielded the best results for all clusterings. The Silhouette was 0.36 (0.29 for K-Means), as a result we present the results of EM in the remainder of the paper. The probability distribution of clusters produced by EM showed that at least 75% of the users have a probability higher than 0.9 to belong to the first assigned cluster.

We inspected those clusters and interpreted them according to the feature distributions.

We observed that five clusters (out of nine) bear very minimal differences in the feature distributions and we could not identify distinct behaviour. As a result, we decided to merge those cluster and assume that correspond to similar user behaviour. The reason for this is the highly skewed data: most users have very low activity and the majority of messages is contributed by a small fraction of users. Complementary we observe a hierarchical structure among clusters (which also explains the aforementioned results) that shows smaller differences among users in the same clusters. The rest of the analysis considers five distinct clusters which are presented in Figure 2.

We examined representative users and their activity in each cluster to confirm the cluster interpretation. Similar user roles were also identified in the literarture [11, 3, 1]. Any differences with state-of-the-art lie in the different features selected and the platform differences. The five user roles that we identified include:

- Stars: This user role includes extremely popular users (e.g. celebrities, athletes). As seen on Figure 2a stars have extremely high number of followers and their are selective in their friends (followees). They are not so active as users in other clusters, but their messages receive many reactions. They are also mentioned very often, mostly because they are famous. In most cases they are verified and have a url in their profile.

- Information Sources: Users in this cluster are news sources and popular users in particular domains, e.g. bloggers. They have a high number of followers but the gap with the friends is not so extreme as in the case of stars. They are extremely active and engaged, but at the same time they trigger many reactions They are also more conversational compared to stars, indicated by the number of replies. They are being mentioned less than the stars and they are not always verified (e.g. bloggers recognised in particular domains).

- Daily chatters: These users are the most prolific writers (compared to all clusters) propagating both original information or retweeting. They are not so popular and recongnised as the previous clusters. They are mainly talking about their daily routines and reproducing information of what is happening around the world.

- Listeners: These accounts contribute rarely, do not receive reactions and have significantly more followees that followers. Note that this cluster is under-represented in this dataset, since users with more that two messages in Olympics 2012 are considered which is already excluding the true Twitter listeners.

- Average user: This category falls in between of daily chatters and listeners. These users are relatively active, receiving some reactions. They have comparable number of followers and friends Amplifiers are also found in this category that receive information and propagate it further. Note that this user role includes five merged clusters and contains the majority of users. This means that the dominant cluster of average user has small variations (hierarchical structure) which are not easily interpretable.

We examined representative users and their activity in each cluster to confirm the cluster interpretation. Similar user roles were also identified in the literarture [11, 3, 1]. Any differences with state-of-the-art lie in the different features selected and the differences of social media platforms evaluated.

6. STRUCTURAL POSITIONS

After reconstructing information cascades and identifying prominent user roles, we can correlate these two aspects by investigating which positions different user roles occupy in information cascades.

For that, we need to define and compute metrics on information cascades that reveal structural positions for each user; Such structural metrics reveal the influence exerted by users and their centrality in information cascade graphs. We compute the following metrics that show influence and centrality in the cascade graphs for each user:

- *(shortest) Distance to the root* shows whether particular nodes are roots or close to the root, which means that they are influential or have fast access to information.

- *(shortest) Distance to the leaves* reveal nodes who do not trigger significant further reactions.

- *Closeness centrality* measures the distance from a node to all other nodes which demonstrates how central a node in the graph is.

- *Betweeness centrality* measures the number of shortest paths that pass through a node and reveals the amount of information flow that a node controls.

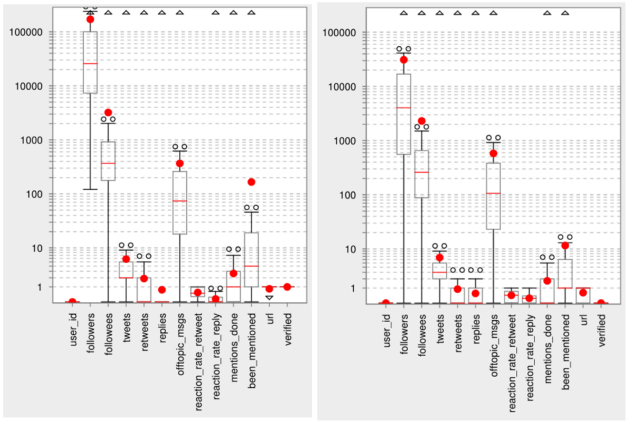
- *Root influence* measures the fraction of nodes who reacted directly to the root and reveals how influential the root is compared to other nodes in the cascade.

- *Indegree* shows the number of different influencers or the amount of influence a node needs to react to incoming information.

- *Outdegree* shows how many nodes a particular node influences.

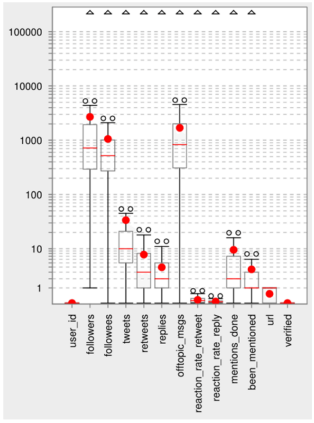
Next, we compute these metrics for the different user roles that were identified in Section 5. We assume that different user roles will demosntarte considerable differenec in terms of their influence and centrality in information cascade graphs.

In order to model behaviours with regard to the the aforementioned metrics for separate user roles, we associate each of the metrics with intensity levels (e.g. low, medium, high), according to the range of their distribution.We split the observations of every metric in three equizised quantiles (0-33,3% for low, 33,3-66 % ,6 for medium and 66,6-100 % for high) which facilitates the comparison of different user roles with regard to these metrics. A similar approach was followed by [20]. By doing this we can

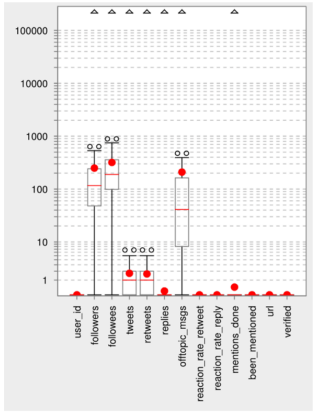


(a) Stars

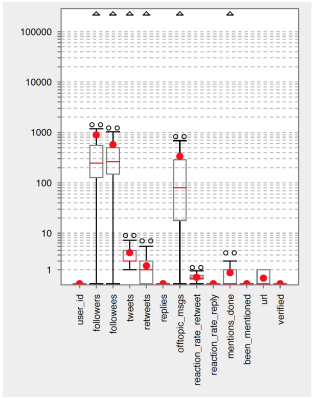
(b) Information sources



(c) Daily Chatters



(d) Listeners



(e) Average User

Figure 2: User Roles

answer questions like: Do stars have a high outdegree compared to the other user roles?

We form similar questions - hypotheses that are going to be confirmed or rejected according to evidence from the data.

The Hypotheses that were successfully confirmed include:

- Stars are creating original content and they demonstrate overall low indegree
- Stars are influencing a lot of others and they have high outdegree
- Since stars are influencing many others directly, they should also be "close" to them in the graph and demonstrate high closeness centrality
- Stars are often observed at the cascade root
- When stars are at the root, they have high root influence; other users are reacting mostly because they are famous
- Daily Chatters and Listeners are positioned at the leaves because they fail to trigger further reactions
- Daily Chatters and Listeners are not central in the cascade graph and have low betweeness centrality

We did not collect enough evidence that positions daily chatters and listeners in the periphery of the graph by demonstrating low closeness centrality. In reality, it is often the case that daily chatters are influenced directly from the root because of their fast reactions, which also brings them "closer" to other nodes.

For information sources we failed to confirm any hypotheses, since this user role acts either as root, or can be found within diffusion paths. Given their diverse behaviour of being popular but at the same time being engaged with others, they can occupy multiple positions over the information cascades.

We also failed to confirm any hypothesis for the average user: these users can take multiple positions on the cascades either in the middle as amplifiers or at the leaves.

The aforementioned Hypotheses were tested by the Mann-Whitney-Wilcoxon test, which is not parametric. The hypotheses that were confirmed, were statistically significant on the 0.001 level. We tested the differences in means for the users in each user role versus the full user population and the differences in means according to the 3 quantiles.

In general, we can observe that user roles at the end of the spectrum (stars, listeners and daily chatters) are correlated with cascade structure. The user roles of information sources and average user needs further investigation in terms of their behaviour, since these users occupy multiple positions in the cascades. Also we need further evaluations to understand the subtle differences of daily chatters and listeners into the information cascades. These two user roles seem to have very different behavioural patterns but they occupy similar structural positions. In order to validate the importance of such analysis, we will further evaluate such hypotheses in larger datasets and more social media platforms.

7. CONCLUSION AND FUTURE WORK

In this paper we have presented a study that correlates user roles with structural aspects of information diffusion in Twitter. While we identified particular user roles that correlate with the cascade structure, this work has some limitations. For the user roles that constitute the core of social media (average user and information sources) we failed to confirm any hypotheses and we need to

investigate further their online behaviour. For future work, we plan to identify information cascade shapes (stars, complex structures, long paths, etc) and correlate such shapes with user roles. This analysis will help us to gain a better understanding into human interactions and influence in social media and provide valuable insights for information diffusion.

8. REFERENCES

[1] S. Angeletou, M. Rowe, and H. Alani. Modelling and analysis of user behaviour in online communities. In *International Semantic Web Conference*, pages 35–50. Springer, 2011.

[2] J. Chan, C. Hayes, and E. M. Daly. Decomposing discussion forums and boards using user roles. 2010.

[3] R. Edgar, P. F. Alexandre, P. Caladoa, and H. Sofia-Pinto. User profiling on twitter. *Semantic Web Journal*, 2011.

[4] M. Farajtabar, M. Gomez-Rodriguez, Y. Wang, S. Li, H. Zha, and L. Song. Co-evolutionary dynamics of information diffusion and network structure. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 619–620, 2015.

[5] M. Goetz, J. Leskovec, M. McGlohon, and C. Faloutsos. Modeling blog dynamics. In *ICWSM*, 2009.

[6] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 1019–1028, 2010.

[7] M. Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420, 1978.

[8] A. Guille, H. Hacid, C. Favre, and D. A. Zighed. Information diffusion in online social networks: a survey. *SIGMOD Record*, 42:17–28, 2013.

[9] H. W. Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[10] C. Hui, Y. Tyshchuk, W. A. Wallace, M. Magdon-Ismail, and M. Goldberg. Information cascades in social media in response to a crisis: A preliminary model and a case study. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 653–656, 2012.

[11] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[12] K. Y. Kamath, J. Caverlee, K. Lee, and Z. Cheng. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 667–678, 2013.

[13] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600, 2010.

[14] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), 2007.

[15] J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *SDM*, volume 7, pages 551–556, 2007.

[16] C.-T. Li, Y.-J. Lin, and M.-Y. Yeh. The roles of network communities in social information diffusion. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 391–400. IEEE, 2015.

[17] S. A. Myers and J. Leskovec. The bursty dynamics of the twitter information network. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 913–924, 2014.

[18] M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[19] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[20] M. Rowe, M. Fernandez, S. Angeletou, and H. Alani. Community analysis through semantic rules and role composition derivation. *Web Semantics: Science, Services and Agents on the World Wide Web*, 18(1):31–47, 2013.

[21] I. Taxidou and P. M. Fischer. Online analysis of information diffusion in twitter. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '14 Companion, 2014.

[22] R. Tinati, L. Carr, W. Hall, and J. Bentwood. Identifying communicator roles in twitter. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 1161–1168, 2012.

[23] Y. Yang, J. Tang, C. W.-k. Leung, Y. Sun, Q. Chen, J. Li, and Q. Yang. Rain: Social role-aware information diffusion. In *AAAI*, pages 367–373, 2015.

[24] Z. Zhou, R. Bandari, J. Kong, H. Qian, and V. Roychowdhury. Information resonance on twitter: Watching iran. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 123–131, 2010.

[25] B. Zong, Y. Wu, A. K. Singh, and X. Yan. Inferring the underlying structure of information cascades. In *2012 IEEE 12th International Conference on Data Mining*, pages 1218–1223. IEEE, 2012.