Going Beyond Structural Communities in Online Social Networks

David Darmon, Elisa Omodei, Joshua Garland

Community detection in online social networks is typically based on the analysis of the explicit connections between users, such as "friends" on Facebook and "followers" on Twitter. But online users often have hundreds or even thousands of such connections, and many of these connections do not correspond to real friendships or more generally to accounts that users interact with. We claim that community detection in online social networks should be question-oriented and rely on additional information beyond the simple structure of the network. The concept of 'community' is very general, and different questions such as "who do we interact with?" and "with whom do we share similar interests?" can lead to the discovery of different social groups. We propose looking at when and how communities motivated by these different questions overlap, and whether different approaches to asking the question, "What community are you in?" leads to different insights about a social network. For example, a user on Twitter might connect mostly with computational social scientists, talk mostly about machine learning, interact solely with close friends (who may or may not be computational social scientists), and utilize the service only on nights and weekends. Each of these different 'profiles' of the user highlight different views of the user's social network, and represent different types of communities. We divide our approaches into four categories based on the questions outlined above: structure-based, activity-based, topic-based, and interaction-based. The structure-based approach, is most common, and for our data relies on reported follower relationships. To detect the other kinds of communities we weight the structural edges with different scores that aim to measure the different kinds of interactions. For the activity-based communities, we consider only the timing of each user's tweets and ignore any additional content. From this starting point, we can view the behavior of a user u on Twitter as a point process, where at any instant t the user has either emitted a tweet ($X_t(u) = 1$) or remained silent ($X_t(u) = 0$). Thus, we reduce all of the information generated by a user on Twitter to a time series $\{X_t(u)\}$. We can then compute the flow of information between two users u and v by computing the transfer entropy between their time series $X_t(u)$ and $X_t(v)$. For the interaction-based communities, we weight the user-follower network with a measure proportional to the number of mentions and/or retweets between users. Lastly, in order to detect the topical communities, we weight the edges of the user-follower network through a measure based on the number of common hashtags between pairs of users. Hashtags are a good proxy for a tweet's content as hashtags are explicitly meant to be keywords indicating the topic of the tweet. We then used the OSLOM algorithm [6] to detect overlapping communities present in the four weighted, directed networks. We then compared with several different methods the communities obtained in the different weightings, in particular we considered the normalized mutual information. We see that similarity between the coverings is dictated by the generic community type (structural, activity-based, etc.). That is, the transfer entropy coverings are more similar to each other than to any of the other coverings, with a similar result for the mention, retweet, and mention-retweet coverings. Interestingly, the coverings resulting from the different weightings are all more similar to each other than to the structural covering from the unweighted network. Also the covering based on the hashtag similarities are different from all of the other weight-based coverings. We also explored the results qualitatively and found topic-based communities in which most members are not found in the same community in the other networks, indicating that while these people talk about the same things and can therefore be considered a community based on their content, they do not strongly interact with each other nor behave the same, and so belong to different social groups with respect to interactions and behavior. Another interesting example is a community whose topics tend to focus on Colorado. These users do not belong to the same community in the interaction-based network, but most of them do belong to the same community in the activity-based network. This indicates that these users react to the same events and issues regarding Colorado and are therefore strongly connected in the topic-based and activity-based networks, but at the same time they do not directly interact with each other and are therefore more loosely connected in the interaction-based networks, where they belong to different communities. This work demonstrates that asking the proper question and then crafting an appropriate weighting scheme to answer that question is an unavoidable first step for community detection in online social media, and that, without a clear definition of community, many rich and interesting communities present in online social networks remain invisible.

References

[1] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *ICWSM*, 2011.

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

[3] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: An analysis of a microblogging community. In *Advances in Web Mining and Web Usage Analysis*, pages 118–138. Springer, 2009.

[4] Anne Kao, William Ferng, Stephen Poteet, Lesley Quach, and Rod Tjoelker. Talison-tensor analysis of social media data. In *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*, pages 137–142. IEEE, 2013.

[5] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

[6] Andrea Lancichinetti, Filippo Radicchi, José J. Ramasco, and Santo Fortunato. Finding statistically significant communities in networks. *PLoS ONE*, 6(4):e18961, 04 2011.

[7] Greg Ver Steeg and Aram Galstyan. Information transfer in social media. In *Proc. 21st Int'l World Wide Web Conf.*, pages 509–518. ACM, 2012.

[8] Zhijun Yin, Liangliang Cao, Quanquan Gu, and Jiawei Han. Latent community topic analysis: Integration of community discovery with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):63, 2012.