Question-Oriented Community Detection 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. This so-called *structure-based* approach relies solely on the reported relationships between users—completely ignoring the users activity, content, or interactions. However, users often have hundreds or even thousands of such structural-connections, many of which do not correspond to real friendships or even accounts the user interacts with. We claim that community detection in online social networks should be *question-oriented* and rely on additional information beyond the simple (often spurious) structure of the network. The concept of ‘community’ is very general, and we have found that different questions such as “who does a user interact with?”, “from whom does a user receive information?” and “with whom does a user share similar interests?” can lead to the discovery of many rich *and* interesting communities that the structure-based analysis alone is completely blind to. 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. We argue that 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, outlined above, relies solely on the known follower relationships, and is used here purely for comparison with the standard. To detect the *question-oriented* communities we weight the structural edges with different scores derived to both answer and give insight into the question posed about the network. For the *activity-based* communities, we consider only the timing of each user’s tweets and ignore any additional content, i.e., we 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 ($X_t(u) = 0$)—reducing all of the information generated by a user to a symbolic time series: $\{X_t(u)\}$. We then compute the directed flow of information between two users u and v by computing the *transfer entropy* between their time series $X_t(u)$ and $X_t(v)$; an information theoretic approach which gives insights into how information transmits in a network as well as insights into who the influential users are. For the *interaction-based* communities, we weight the structural 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 structural 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. It should be noted that a user may be a member of multiple communities simultaneously in a given network, e.g., a user may tweet about “Complexity” *and* “Lucca, Italy”, which may be two separate topical communities, but may overlap for example when a collection of users tweet about ECCS’14. For this reason we seek overlapping communities—often called *coverings*.

To detect covering of each weighted, directed network we use the OSLOM algorithm [1]. We then applied several statistical measures (e.g., mutual information) to compare the communities derived from the different questions with intriguing and informative results. Not surprisingly, there was similarity between coverings dictated by each generic community type (structural, activity-based, etc.). Interestingly, the coverings resulting from the *question-oriented* weightings were all more similar to each other than to the covering from the *structure-based* analysis. We found that the coverings derived from the *topic-based* questions (hashtag weightings) were significantly different from all of the other question-oriented coverings.

We explored these coverings qualitatively and found interesting insight into these networks. For example, we found topic-based communities containing members that were otherwise disjoint based on interaction or activity. These indicate users who tweet about the same content and should therefore be considered a topical community, but do not strongly interact with each other nor use twitter to transmit information to each other, and should therefore belong to different communities with respect to interactions and activity. 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. These are very nice empirical examples that illustrate our main hypothesis: the *questions* you ask about the network dictate the communities which are discoverable and it is insufficient to simply do a single (weighted or unweighted) community detection, to fully understand a networks community structure several weightings should be derived from interesting questions about the network and then the coverings should be compared qualitatively.

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