Followers are Not Enough: Beyond Structural Communities in Online Social Networks

Abstract

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. In this paper we focus on three types of communities beyond structural communities: activity-based, topic-based, and interaction-based. We analyze a Twitter dataset using three different weightings of the structural network meant to highlight these three community types, and then infer the communities associated with these weightings. We show that the communities obtained in the three weighted cases are highly different from each other, and from the communities obtained by considering only the unweighted structural network. Our results confirm that asking a precise question is an unavoidable first step in community detection in online social networks, and that different questions can lead to different insights into the network under study.

1 Introduction

Networks play a central role in online social networks like Twitter, Facebook, and Google+. These services allow a user to interact with other based on the online social network they curate, also known as contact filtering (Cazabet, Leguistin, and Amblard 2012). For example, 'friends' on Facebook represent reciprocal links for sharing information, while 'followers' on Twitter allow a single user to broadcast information in a one-to-many fashion. Central to all of these interactions is the fact that the *structure* of the social network dictates the ways that information can be broadcast or diffuse through the service.

Because of the importance of structural networks, a large amount of work has focused on using structural networks to detect communities in online social media. Here, we use the term community to mean the standard definition of collection of nodes (users) within the network who are more highly connected to each than to users outside of the community (Newman and Girvan 2004). For instance, in (Java et al. 2009), the authors use a friendship network to determine communities within Twitter, and note that conversations tend to occur within these communities. The approach of focusing on structural networks makes sense for 'real-world' sociological experiments, where obtaining additional information about users interactions may be expensive and time-consuming. However, with the prevalence of large, rich data sets for online social networks, additional information beyond the structure alone may be incorporated, and more realistically reflects how users interact with each other on social media services (Grabowicz et al. 2012).

A large body of work exists on methods for automatic detection of communities within networks (Newman 2004; Newman and Girvan 2004; Rosvall and Bergstrom 2008; Blondel et al. 2008; Lancichinetti et al. 2011). See (Fortunato 2010) for a recent review. All these methods *begin* with a given network, and then attempt to uncover structure present in the network. That is, they are agnostic to how the network was constructed. Because of this, we propose that in answering any sort of question about the communities present in a data set, it is important to begin with a clear picture of the community type under consideration, and then to tailor the construction of the network and the use of detection algorithms towards that goal.

This is especially true for social network analysis. In online social networks, 'community' could could refer to several possible structures. The simplest definition of community, as we have seen, might stem from the network of explicit connections between users on a service (friends, followers, etc.). On small time scales, these connections are more or less static, and we might instead determine communities based on who is talking to whom. On a more abstract level, a user might consider themselves part of a community of people discussing similar topics. We might also define communities as collections of people who exhibit similar behaviors on a service, as in a communities of teenagers vs. elderly users. We can characterize these types of com-

munities based on the types of questions we might ask about them:

- **Structure-based:** Who are you friends with? Who do you follow?

- **Activity-based:** Who do you act like?

- **Topic-based:** What do you talk about?

- **Interaction-based:** Who do you talk to? Who do you talk about?

We propose looking at when and how communities motivated by 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 soley 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, as outlined above, is most common, and for our data relies on reported follower relationships.

The activity-based approach is motivated by the question of which individuals act in a concerted manner. The main tools for answering this question stem from information theory. We consider each user on an online social network as an information processing unit, but ignore the content of their messages. This viewpoint has been successfully applied to gain insight into local behavior in online social networks (Darmon et al. 2013b). The information processing framework applies equally well to spatially extended systems. In particular, our current activity-based approach was originally motivated by a methodology used to detect functional communities within populations of neurons (Shalizi, Camperi, and Klinkner 2007). This approach has been extended to social systems, detecting communities on Twitter based on undirected information flow (Darmon et al. 2013a). Others have successfully applied a similar viewpoint using transfer entropy, a measure of directed information flow, to perform link detection on Twitter (Ver Steeg and Galstyan 2012). However, transfer entropy has not been used for community detection, and we extend this previous work to incorporate transfer entropy-based communities.

Our topic-based and interaction-based approaches, in contrast to the activity-based approach, rely on the *content* of a user's interaction and ignore its temporal component. The content contains a great deal of information about the communication between users. For example, a popular approach to analyzing social media data is to use Latent Dirichlet Allocation (LDA) to infer topics based on the prevalence of words within a status (Zhao et al. 2011; Michelson and Macskassy 2010). The LDA model can then be used to infer distributions over latent topics, and the similarity of two users with respect to topics may be defined in terms of the distance between their associated topic distributions. Because our focus is not on topic identification, we apply a simpler approach using hashtags as a proxy for topics (Becker, Naaman, and Gravano 2011; Tsur and Rappoport 2012). We can then define the similarity of two users in terms of their hashtags, and use this similarity to build a topic-based network.

The interaction-based approach relies on meta-data about who a user converses with on the social media service. On Twitter, we can use mentions (indicating a directed communication) and retweets (indicating endorsement of another user) to identify conversation. Moreover, we can define a directed influence between two users by considering the attention paid to that user compared to all other users. This allows us to generate a network based on conversations and user interactions.

The activity-based, topic-based, and interaction-based networks allow us to build a more complete picture of the *latent* social network present in online social media, as opposed to the *explicit* social network indicated by structural links. In this paper, we explore the relation between these various possible networks and their corresponding communities. We begin by describing the methodologies used to generate the various types of networks, and infer their community structure. We then explore how the communities of users differ depending on the type of network used. **TK: Put a blurb here about the qualitative study.** Finally, we explore how communication patterns differ across and within the different community types.

2 Related Work

Most previous work on communities in online social networks have focused on detecting communities based on a single type of network. For example, an early paper considered the communities, and associated statistics, inferred from a friendship network in Twitter (Java et al. 2009). More recent work has considered the dynamics of communities based on structural links in Facebook (Nguyen et al. 2011).

Information theoretic, activity-based approaches have been applied previously to the analysis of networks arising in online social media (Ver Steeg and Galstyan 2012; Darmon et al. 2013a), but to the best of our knowledge this is the first use of transfer entropy for community detection.

For interaction-based networks, (Conover et al. 2011) considered both mention and retweet networks in isolation for a collection of users chosen for their political orientation. In (Deitrick and Hu 2013), the authors construct a dynamic network based on simple time-windowed counts of mentions and retweets, and use the evolution of this network to aid in community detection.

There are two broad approaches to topic-based communities in the literature. (Rossi and Magnani 2012) used a set of users collected based on their use of a single hashtag, and tracked the formation of follower and friendship links within that set of users. In (Lim and Datta 2012a), the authors chose a set of topics to explore, and then seeded a network from a celebrity chosen to exemplify a particular topic. Both approaches thus begin with a particular topic in mind, and perform the data collection accordingly. Other approaches

use probabilistic models for the topics that treat community membership as a latent variable (Yin et al. 2012).

A notable exception to the analysis of isolated types of networks is (Lim and Datta 2012b), which considered both structure-based and interaction-based communities on Twitter. However, this study focuses on data collected based on particular topics (country music, tennis, and basketball), and not on a generic subpopulation of Twitter users. Moreover, it did not explore the differences in community structure resulting from the different network weightings, and focuses on aggregate statistics (community size, network statistics, etc.). Another notable exception is (Kao et al. 2013), where the authors use a tensor representation of user data to incorporate retweet and hashtag information into a study of the social media coverage of the Occupy Movement. The tensor can then be decomposed into factors in a generalization of the singular value decomposition of a matrix, and these factors can be used to determine 'salient' users. However, this approach focused on data for a particular topic (the Occupy Movement) and did not collect users based on a structural network.

3 Methodology

3.1 Community Detection

In network theory communities are usually defined as group of nodes more densely connected among each other than with the rest of the network. Community detection is a well studied but not yet solved problem, and several different methods and algorithms have been proposed. For a complete review of the subject we refer the reader to (Fortunato 2010). In our study we deal with *weighted directed* networks, and we are interested in finding *overlapping* modules, rather than partitions of the network, since people can belong to different social groups: they can be interacting with different groups of people (their college friends, their co-workers, their family, etc), and they can belong to different topical communities (a person can be interested both in cycling and politics and talk about the two topics with different groups of people). Most community detection algorithms developed so far are built to find partitions and only very few to find overlapping communities (Baumes et al. 2005; Palla et al. 2005; Zhang, Wang, and Zhang 2007; Gregory 2007; Nepusz et al. 2008; Lancichinetti, Fortunato, and Kertész 2009a; Evans and Lambiotte 2009; Kovács et al. 2010) Among these, even fewer can also deal with directed or weighted networks. Palla et al. clique percolation method can for example take into account these two features, but not both at the same time (Palla et al. 2005). Lancichinetti et. al recently proposed the first method that is able to deal with all these features and a few more at the same time, called OSLOM (Order Statistics Local Optimization Method) (Lancichinetti et al. 2011). Their method "is based on the local optimization of a fitness function expressing the statistical significance of clusters with respect to random fluctuations" (**taken from their paper, to rephrase**), and it allowed us to detect overlapping communities on our weighted directed networks.

3.2 The Dataset

The data consists of the Twitter statuses of 15000 users over a 9 week period (from April 25th to June 25th 2011). The users are embedded in a network collected by performing a breadth-first expansion from a seed user. Once the seed user was chosen, the network was expanded to include his/her followers, only including users considered to be active (users who tweeted at least once per day over the past one hundred tweets). Network collection continued in this fashion by considering the active followers of the active followers of the seed, and so on. Since one of the kind of communities we want to explore is based on the explicit interactions between users, based on retweets and mentions, we filter the dataset in order to take into account only users who make use of this kind of features in their tweets. We define en event of outgoing information for a given user u as either a mention made by u of another user in the network, or a retweet by another user of one of u's tweets. When we mention someone we are in fact sending him some information, and when we are retweeted it means that the person that retweeted us has received some information from us and is sharing it. Symmetrically, we define an event of incoming information for u as either being mentioned, or retweeting another user. We filtered the network by eliminating all the users that have in their tweeting history less than 9 outgoing information and 9 incoming information events, i.e. less than one event per type per week on average. We then further restricted our analysis to the strong giant connected component of the unweighted directed network built from the filtered set of users and whose link represent a user-follower relationship. In this study the link is directed from the user to the follower because this is the direction in which the information flows. The final network consists of 6917 nodes and 1481131 edges.

Show the distribution of the number of mentions / retweets / haghstags / tweets / followers per user ?

3.3 Activity-Based Communities and Transfer Entropy

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 ($X_t(u) = 0$). This is the view of a user's dynamics taken in (Ver Steeg and Galstyan 2012) and (Darmon et al. 2013b). Thus, we reduce all of the information generated by a user on Twitter to a time series $\{X_t(u)\}$ where t ranges over the time interval for which we have data (9 weeks in this case). Because status updates are only collected in discrete, 1-second time intervals, it is natural to consider the only the discrete times $t = 1\,\mathrm{s}, 2\,\mathrm{s}, \ldots$, relative to a reference time. 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 $X_t(v)$.

Let $\{X_t\}$ and $\{Y_t\}$ be two strong-sense stationary stochastic processes. We use the notation X_{t-k}^t to denote the values of the stochastic process from time $t-k$ to time t, $X_{t-k}^t = (X_{t-k}, X_{t-(k-1)}, \ldots, X_{t-1}, X_t)$. The lag-$k$ trans-

fer entropy of Y on X is defined as

$$\mathrm{TE}^{(k)}_{Y \to X} = H\left[X_t | X_{t-k}^{t-1}\right] - H\left[X_t | X_{t-k}^{t-1}, Y_{t-k}^{t-1}\right], \quad (1)$$

where

$$H\left[X_t | X_{t-k}^{t-1}\right] = -E\left[\log_2 p(X_t | X_{t-k}^{t-1})\right] \quad (2)$$

and

$$H\left[X_t | X_{t-k}^{t-1}, Y_{t-k}^{t-1}\right] = -E\left[\log_2 p(X_t | X_{t-k}^{t-1}, Y_{t-k}^{t-1})\right] \quad (3)$$

are the usual conditional entropies over the conditional (predictive) distributions $p(x_t | x_{t-k}^{t-1})$ and $p(x_t | x_{t-k}^{t-1}, y_{t-k}^{t-1})$. This formulation was originally developed in (Schreiber 2000), where transfer entropy was proposed as an information theoretic measure of *directed* information flow. Formally, recalling that $H\left[X_t | X_{t-k}^{t-1}\right]$ is the uncertainty in X_t given its values at the previous k time points, and that $H\left[X_t | X_{t-k}^{t-1}, Y_{t-k}^{t-1}\right]$ is the uncertainty in X_t given the joint process $\{(X_t, Y_t)\}$ at the previous k time points, transfer entropy measures the reduction in uncertainty of X_t by including information about Y_{t-k}^{t-1}, controlling for the information in X_{t-k}^{t-1}. By the 'conditioning reduces entropy' result (Cover and Thomas 2012)

$$H[X|Y,Z] \leq H[X|Y], \quad (4)$$

we can see that transfer entropy is always non-negative, and is zero precisely when $H\left[X_t | X_{t-k}^{t-1}\right] = H\left[X_t | X_{t-k}^{t-1}, Y_{t-k}^{t-1}\right]$, in which case knowing the past k lags of Y_t does not reduce the uncertainty in X_t. If the transfer entropy is positive, then $\{Y_t\}$ is considered causal for $\{X_t\}$ in the Granger sense (Granger 1963).

In estimating the transfer entropy from finite data, we will assume that the process (X_t, Y_t) is jointly stationary, which gives us that

$$p(x_t | x_{t-k}^{t-1}) = p(x_{k+1} | x_1^k) \quad (5)$$

and

$$p(x_t | x_{t-k}^{t-1}, y_{t-k}^{t-1}) = p(x_{k+1} | x_1^k, y_1^k) \quad (6)$$

for all t. That is, the predictive distribution only depends on the past, not on when the past is observed[1]. Given this assumption, we compute estimators for $p(x_{k+1} | x_1^k)$ and $p(x_{k+1} | x_1^k, y_1^k)$ by 'counting': for each possible past (x_1^k, y_1^k), we count the number of times a future of type x_{k+1} occurs, and normalize. Call these estimators $\hat{p}(x_{k+1} | x_1^k)$ and $\hat{p}(x_{k+1} | x_1^k, y_1^k)$. Then the plug-in estimator for the transfer entropy is

$$\widehat{\mathrm{TE}}^{(k)}_{Y \to X} = \hat{H}\left[X_t | X_{t-k}^{t-1}\right] - \hat{H}\left[X_t | X_{t-k}^{t-1}, Y_{t-k}^{t-1}\right] \quad (7)$$

where we use the plug-in estimators $\hat{H}\left[X_t | X_{t-k}^{t-1}\right]$ and $\hat{H}\left[X_t | X_{t-k}^{t-1}, Y_{t-k}^{t-1}\right]$ for the entropies. It is well known that the plug-in estimator for entropy is biased (Paninski 2003). To account for this bias, we use the Miller-Madow adjustment to the plug-in estimator (Miller 1955).

---

[1]We really only need *conditional* stationarity (Caires and Ferreira 2003), but stationarity implies conditional stationarity

For the communities based on transfer entropy, we weight each edge from a user u to a follower f by the estimated transfer entropy of the user u on f,

$$w^{\mathrm{TE}(k)}_{u \to f} = \widehat{\mathrm{TE}}^{(k)}_{X(u) \to X(f)}. \quad (8)$$

Operationally, we expect users to interact with Twitter on a human time scale, and thus the natural one-second time resolution is too fine. We coarsen each time series by considering non-overlapping time intervals ten minutes in length. For each time interval, we record a 1 if the user has tweeted during that time interval, and a 0 if they have not. Thus, the new coarsened time series now captures whether or not the user has been active on Twitter over any given ten minute time interval in our data set. We then compute the transfer entropy on these coarsened time series taking k to range from 1 to 6, which corresponds to a lag of ten minutes to an hour.

3.4 Interaction-Based Communities and Mention / Retweet Weighting

A way of tracking the flow of information through users is by considering two useful features of this social network: mentions and retweets. Through mentions users can in fact send a direct message to other users. And a retweet means that a piece of information from a user has been captured by a follower and shared with his/her own followers. We can therefore define interaction-based communities by weighting the user-follower network with a measure proportional to the number of mentions and retweets between users. In more details, for each couple of user-follower u and f, we use the arithmetic / harmonic (**which one?**) mean of the two following numbers:

$$\frac{\text{\# mentions of } f \text{ by } u}{\text{\# total mentions of } f} \quad (9)$$

and

$$\frac{\text{\# retweets of } u \text{ by } f}{\text{\# total retweets made by } f}. \quad (10)$$

For the communities based on mention-retweets, we weight each edge from a user u to a follower f by

$$w^{\mathrm{MR}}_{u \to f} = \frac{1}{2}\left(\frac{\text{\# mentions of } f \text{ by } u}{\text{\# total mentions of } f} + \frac{\text{\# retweets of } u \text{ by } f}{\text{\# total retweets made by } f}\right). \quad (11)$$

3.5 Topic-based Communities and Hashtag Weighting

Another kind of community is the one based on the content of the tweets, and relies on the idea of finding people that talk about the same things. In order to detect this kind of communities, we weight the edges of the user-follower network through a measure based on the number of common hashtags between the two users. Hashtags are in fact a good proxy for this, since they are explicitly meant to be key-words indicating a particular topic. Moreover they are widely used and straightforwardly detectable. We characterize each user u by a vector $\vec{h}(u)$ of length equal to the

number of hashtags in the dataset, and whose elements are defined as

$$h_i(u) = n_i(u) * \log \frac{N}{n_i} \qquad (12)$$

where $n_i(u)$ is the frequency of the hashtag i in the set of user u's tweets, N is the total number of users, and n_i is the number of users that have used the hashtag i in their tweets. This adapted term frequency–inverse document frequency (tf-idf) measure (Salton and McGill 1983) captures the importance of a hashtag in the users's tweets through the first factor, but at the same time smooths it through the second factor by giving less importance to hashtags that are too widely used (as $\frac{N}{n_i}$ approaches one, its logarithm approaches zero). For each couple of user linked in the user-follower network, we then compute the cosine similarity of their respective vectors, and assign the obtained value as the weight of the directed edge(s) connecting them. This weight captures the similarity between users in terms of topic discussed in their tweets. Thus, for communities based on hashtag similarity, we weight each edge from a user u to a follower f by

$$w_{u \to f}^{\text{HT}} = \frac{\vec{h}(u) \cdot \vec{h}(f)}{||\vec{h}(u)|| \; ||\vec{h}(f)||}. \qquad (13)$$

4   Results

4.1   Comparing Aggregate Statistics of Community Structure

We begin by examining the overall statistics for the communities inferred by OSLOM using the weightings defined in Sections 3.3, 3.4, and 3.5. The number of communities by community type is given in Table 1. We see that the topic- and interaction-based networks admit the most communities. The activity-based network admits the least number of communities. One advantage of the OSLOM over many other community detection algorithms is that it explicitly accounts for singleton 'communities': those nodes who do not belong to *any* extant communities. This is especially important when a network is collected via a breadth-first search, as in our network, where we begin from a seed node and then branch out. Such a search, once terminated, will result in a collection of nodes on the periphery of the network that may not belong to any community in the core.

The number of singletons by community type is also shown in Table 1. We see that the topic- and interaction-based communities have the most singletons, with the activity-based community dominating this measure. This result for the activity-based community is an artifact of a property of the retweet/mention weighting: many of the users in the data set (**TK: give actual number**) did not interact with each other, and thus all of their edges were zero-weighted, leading to trivial singletons.

Next we consider the distribution of community sizes across the community types. The complementary cumulative distribution of community sizes is given in Figure 1. Note that the axes are plotted on log-scale, and the horizontal axis begins with non-singleton communities. Thus, for a fixed community size c, Figure 1 shows the proportion of communities of size greater than c for each community type.

Table 1: Number of non-singleton communities and singletons by community type: S(tructural), A(ctivity -based), T(opic-based), and I(nteraction-based).

Community Type	# of Communities	# of Singletons
S	201	308
A, Lag 1	101	951
A, Lag 2	99	600
A, Lag 3	106	611
A, Lag 4	105	668
A, Lag 5	107	632
A, Lag 6	106	642
T	289	1064
I	252	2436

The largest communities for the structural, activity-based, topic-based, and interaction-based networks are 198, 358, 338, and 811 respectively.
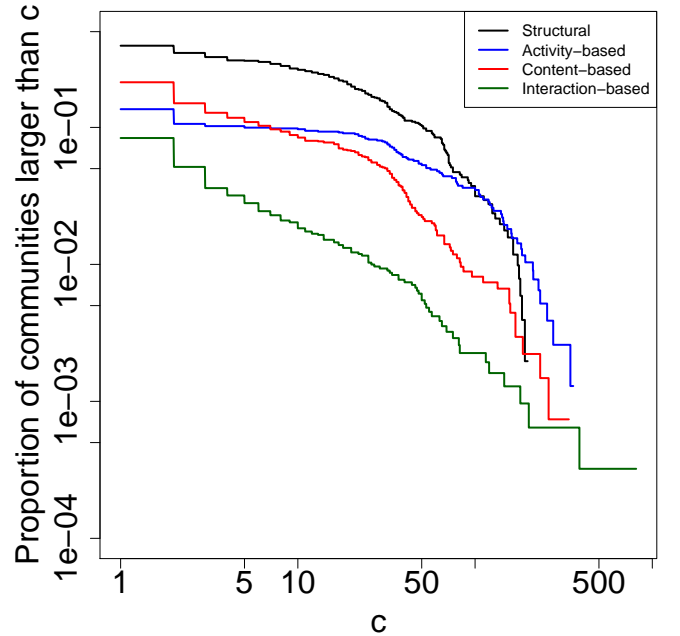

Figure 1: The proportion of communities greater than c in size, across the different community types. Note the logarithmic scale on the horizontal and vertical axes.

Next, we compare the number of users which belong to more than one community. Figure 2 shows the number of users belonging to 2, 3, or 4 communities. We see that as the number of mixed membership communities increases, the number of users with that number of mixed memberships decreases. This is especially true for the activity-based community **TK: speculate on what this means? Or save for the results section?**. In addition, **TK: mention the 5, 6, and 7 cases, not included in the figure**. This corresponded to **TK: investigate which users are the high-overlap and what communities they belong to.**
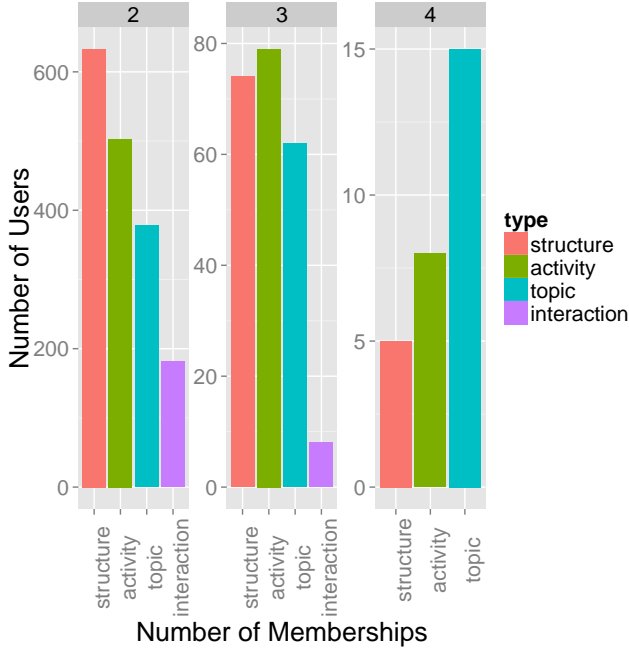
Figure 2: The number of users belonging to 2, 3, or 4 communities, by community type.

4.2 Comparing Community Types with Normalized Mutual Information

Because OSLOM detects *coverings* rather than *partitions* of users, standard cluster comparison methods like variation of information (Meilă 2003) are not appropriate. Instead, we use a generalization of variation of information measure first introduced in (Lancichinetti, Fortunato, and Kertész 2009b), the normalized information. **TK: Etc. Put more here about NMI, what it does, and how to interpret it.**

The normalized mutual information between the various community-types are shown in Figure 3. We see that similarity between the coverings is dictated by the generic covering type, with distinct community structure between the structural, activity-based, and interaction-based communities. Interestingly, the communities resulting from the different weightings are all more similar to each other than to the structural communities from the unweighted network. Also note that the communities based on the hashtag similarities are different from both the activity-based and mention-retweet-based communities.

4.3 Comparing Edges Across Different Community-Types

We have defined communities using four different criteria: structure, activity, interaction, and content. For a fixed community type, edges for a particular community may be partitioned into three sets: those from a node in the community to another node in the community (internal-to-internal), those from a node in the community to a node outside of the community (internal-to-external), and those from a node outside
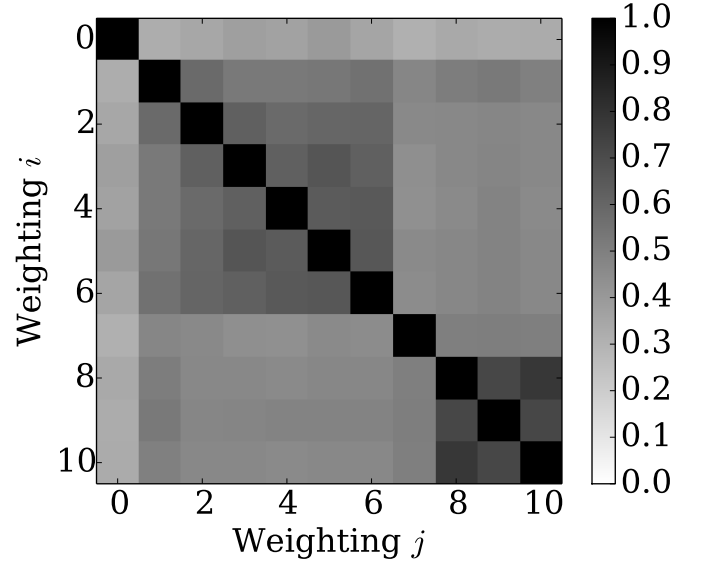

Figure 3: The normalized mutual information between the communities using the different weightings. Weighting 0 corresponds to the structural (binary weighting) network, weightings 1 through 6 correspond to the weighting using the transfer entropy with lag 1 through 6, weighting 7 corresponds to the hashtag similarity, and weightings 8, 9, and 10 correspond to the mention, retweet, and mention-retweet weightings. Values of normalized mutual information close to 1 indicate similarity in the community structure, while values close to 0 indicate dissimilarity. The normalized mutual information is computed with the singletons removed.

the community to a node inside the community (external-to-internal). See Figure 4 for a schematic of this edge partitioning. For a meaningful community, we expect the distribution of weights within the community (internal-to-internal weights) to be different from the distribution of weights without the community (internal-to-external and external-to-internal).

For example, Figure 4 shows the distributions of hashtag-based weights for the largest community defined by the mention-retweet network. We see that the distribution of internal-to-internal hashtag weights has a longer tail than either the external-to-internal or internal-to-external hashtag weights, with edges within the community having higher weights than edges crossing the boundary of the community. Thus, while the community was defined in terms of interactions, we still see a shift in the distribution of *topic-similarity*.

This shift in the distribution was typical of many of the community type / weight pairings. To explore this further, for each of the top 100 largest communities defined by a particular community type (structure-based, behavior-based, content-based, or activity-based), we computed the median internal-to-internal, external-to-internal, and internal-to-external weight for that community for each weight type. These values summarize the typical value of each of these
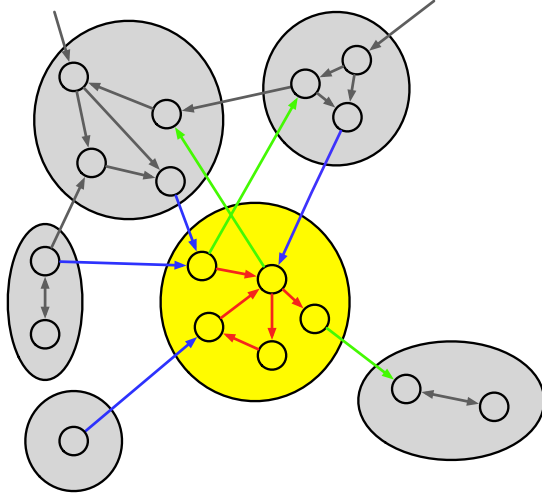
Figure 4: An example of the edges considered in determining the edge weight distribution for a given community (the focal community is in yellow). We focus on the internal-to-internal (red), internal-to-external (green), and external-to-internal (blue) edges. For a given focal community, all other edges (grey) are not considered.
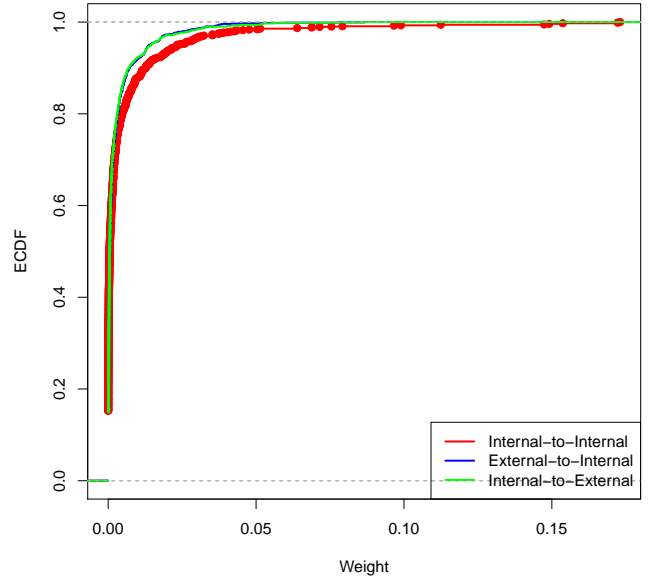

Figure 5: The proportion of edges with a weight at least as large as the weight on the horizontal axis, across the types of edges described in Figure 4. The community is defined by user interactions, and the edge weights are determined by topics.

Table 2: The median value across the 100 largest communities for the ratio of the median internal-to-internal weight to median external/internal-to-internal/external weight for the different community/weight pairings.
* For mention-retweet weights, weight zero edges were excluded from the computation of the median.

		Weight		
		TE	MR	HT
Community	S	0.96/0.94	1.7/2.1*	9.0/8.0
	TE	1.0/0.96	1.5/2.4*	24/17
	MR	0.83/0.86	3.2/4.4	10/8.5
	HT	0.9/0.89	2.4/2.6*	28/26

types of edges. We can then compute the ratio of the median internal-to-internal weight and the median external/internal-to-internal/external weight, which captures the proportional shift in the distribution. Finally, we compute the median of this proportional shift across all 100 of the largest communities. These values are summarized in Table 4.3.

We see that overall, the

5 Discussion

5.1 Questions Influence Community Structure

6 Conclusions

References

Baumes, J.; Goldberg, M. K.; Krishnamoorthy, M. S.; Magdon-Ismail, M.; and Preston, N. 2005. Finding communities by clustering a graph into overlapping subgraphs.

In Guimarães, N., and Isaías, P. T., eds., *IADIS AC*, 97–104. IADIS.

Becker, H.; Naaman, M.; and Gravano, L. 2011. Beyond trending topics: Real-world event identification on twitter. In *ICWSM*.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10):P10008.

Caires, S., and Ferreira, J. 2003. On the nonparametric prediction of conditionally stationary sequences. *Probability, Networks and Algorithms* (4):1–32.

Cazabet, R.; Leguistin, M.; and Amblard, F. 2012. Automated community detection on social networks: useful? efficient? asking the users. In *Proceedings of the 4th International Workshop on Web Intelligence & Communities*, 6. ACM.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Menczer, F.; and Flammini, A. 2011. Political polarization on twitter. In *ICWSM*.

Cover, T. M., and Thomas, J. A. 2012. *Elements of information theory*. John Wiley & Sons.

Darmon, D.; Omodei, E.; Flores, C. O.; Seoane, L. F.; Stadler, K.; Wright, J.; Garland, J.; and Barnett, N. 2013a. Detecting communities using information flow in social net-

works. In *Proceedings of the Complex Systems Summer School*. Santa Fe Institute.

Darmon, D.; Sylvester, J.; Girvan, M.; and Rand, W. 2013b. Understanding the predictive power of computational mechanics and echo state networks in social media. *HUMAN* 2(1):pp–13.

Deitrick, W., and Hu, W. 2013. Mutually enhancing community detection and sentiment analysis on twitter networks. *Journal of Data Analysis and Information Processing* 1:19.

Evans, T. S., and Lambiotte, R. 2009. Line graphs, link partitions, and overlapping communities. *Phys. Rev. E* 80:016105.

Fortunato, S. 2010. Community detection in graphs. *Physics Reports* 486(3):75–174.

Grabowicz, P. A.; Ramasco, J. J.; Moro, E.; Pujol, J. M.; and Eguiluz, V. M. 2012. Social features of online networks: The strength of intermediary ties in online social media. *PLOS ONE* 7(1):e29358.

Granger, C. W. J. 1963. Economic processes involving feedback. *Information and Control* 6(1):28–48.

Gregory, S. 2007. An algorithm to find overlapping community structure in networks. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2007)*, 91–102. Springer-Verlag.

Java, A.; Song, X.; Finin, T.; and Tseng, B. 2009. Why we twitter: An analysis of a microblogging community. In *Advances in Web Mining and Web Usage Analysis*. Springer. 118–138.

Kao, A.; Ferng, W.; Poteet, S.; Quach, L.; and Tjoelker, R. 2013. Talison-tensor analysis of social media data. In *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*, 137–142. IEEE.

Kovács, I. A.; Palotai, R.; Szalay, M. S.; and Csermely, P. 2010. Community landscapes: An integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. *PLoS ONE* 5(9):e12528.

Lancichinetti, A.; Radicchi, F.; Ramasco, J. J.; and Fortunato, S. 2011. Finding statistically significant communities in networks. *PLoS ONE* 6(4):e18961.

Lancichinetti, A.; Fortunato, S.; and Kertész, J. 2009a. Detecting the overlapping and hierarchical community structure in complex networks. *New J. Phys.* 11(3).

Lancichinetti, A.; Fortunato, S.; and Kertész, J. 2009b. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics* 11(3):033015.

Lim, K. H., and Datta, A. 2012a. Following the follower: Detecting communities with common interests on twitter. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, 317–318. ACM.

Lim, K. H., and Datta, A. 2012b. Tweets beget propinquity: Detecting highly interactive communities on twitter using tweeting links. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*, volume 1, 214–221. IEEE.

Meilă, M. 2003. Comparing clusterings by the variation of information. In *Learning theory and kernel machines*. Springer. 173–187.

Michelson, M., and Macskassy, S. A. 2010. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, 73–80. ACM.

Miller, G. A. 1955. Note on the bias of information estimates. *Information theory in psychology: Problems and methods* 2:95–100.

Nepusz, T.; Petróczi, A.; Négyessy, L.; and Bazsó, F. 2008. Fuzzy communities and the concept of bridgeness in complex networks. *Phys. Rev. E* 77:016107.

Newman, M. E., and Girvan, M. 2004. Finding and evaluating community structure in networks. *Physical review E* 69(2):026113.

Newman, M. E. 2004. Fast algorithm for detecting community structure in networks. *Physical review E* 69(6):066133.

Nguyen, N. P.; Dinh, T. N.; Xuan, Y.; and Thai, M. T. 2011. Adaptive algorithms for detecting community structure in dynamic social networks. In *INFOCOM, 2011 Proceedings IEEE*, 2282–2290. IEEE.

Palla, G.; Derényi, I.; Farkas, I.; and Vicsek, T. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435(7043):814–818.

Paninski, L. 2003. Estimation of entropy and mutual information. *Neural Computation* 15(6):1191–1253.

Rossi, L., and Magnani, M. 2012. Conversation practices and network structure in twitter. In *ICWSM*.

Rosvall, M., and Bergstrom, C. T. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 105(4):1118–1123.

Salton, G., and McGill, M. J. 1983. *Introduction to modern information retrieval*. New York: McGraw-Hill.

Schreiber, T. 2000. Measuring information transfer. *Physical review letters* 85(2):461.

Shalizi, C. R.; Camperi, M. F.; and Klinkner, K. L. 2007. Discovering functional communities in dynamical networks. In *Statistical network analysis: Models, issues, and new directions*. Springer. 140–157.

Tsur, O., and Rappoport, A. 2012. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 643–652. ACM.

Ver Steeg, G., and Galstyan, A. 2012. Information transfer in social media. In *Proc. 21st Int'l World Wide Web Conf.*, 509–518. ACM.

Yin, Z.; Cao, L.; Gu, Q.; and Han, J. 2012. Latent community topic analysis: Integration of community discovery with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3(4):63.

Zhang, S.; Wang, R.-S.; and Zhang, X.-S. 2007. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A: Statistical Mechanics and its Applications* 374(1):483–490.

Zhao, W. X.; Jiang, J.; Weng, J.; He, J.; Lim, E.-P.; Yan, H.; and Li, X. 2011. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*. Springer. 338–349.