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

Abstract

[[Joshua]]
[[David]]
[[Elisa]]

Introduction

Networks play a central role in online social networks like Twitter, Facebook, and Google+. These services have become popular largely because of the fact that they allow for the interaction of millions of users in ways restricted by the network each user chooses to be part of. Central to this interaction is the concept

A large body of work exists on methods for automatic detection of communities within networks **(lots of references go here, starting with Newman and ending with Fortunato)**. 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, 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 type of community under consideration, and then to tailor the construction of the network and the use of detection algorithms

This is especially true for social network analysis. In online social networks, 'community' could mean many things. The simplest definition of community 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 discuss 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 communities based on the types of questions we might ask about them:

- **Structure-based:** Who are you friends with?
- **Interaction-based:** Who do you talk to?
- **Topic-based:** What do you talk about?

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

Most previous work on communities in online social networks have focused on these types of communities in isolation. For instance **(References from links.txt go here.)**

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, **(Put a 'worked example' here as to how a single person might belong to various different types of communities, and how these types of communities might reveal different insights.)**

In particular, we can break down our approaches into two broad categories: content-free and content-full. The content-free 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 **(Cite Shannon here, with his ideas that the content of the message doesn't matter for information theory?)**. 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 content-free 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).

In contrast to the content-free approach, a content-full approach would take into account the *content* being transmitted via an online social network. This content has a great deal of social information embedded in it. For example, on Twitter, a tweet might have a hashtag (indicating a topic), a mention or reply (indicating a directed communication), or a retweet (indicating endorsement of another user). This information allows us to build a more complete picture of the *latent* social network, as opposed to the *explicit* social network indicated by friend and follower links.

Many approaches have explicitly accounted for this information in their definition of a community. **(Put references to work done using mentions / replies / retweets / hash tags, etc.)**

Wrap up. Probably write this *after* more of our results are in.

Activity-Based v. Interaction-Based v. Topic-Based Communities

Lorem ipsum.

Methodology

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 (**ref: Fortunato, Santo. "Community detection in graphs." Physics Reports 486.3 (2010): 75-174.**). 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 (**ref: J. Baumes, M.K. Goldberg, M.S. Krishnamoorthy, M.M. Ismail, N. Preston, Finding communities by clustering a graph into overlapping subgraphs, in: N. Guimaraes, P.T. Isaias (Eds.), IADIS AC, IADIS, 2005, pp. 97104; Palla, Gergely, et al. "Uncovering the overlapping community structure of complex networks in nature and society." Nature 435.7043 (2005): 814-818.; S. Zhang, R.-S. Wang, X.-S. Zhang, Identification of overlapping community structure in complex networks using fuzzy c-means clustering, Physica A 374 (2007) 483490; 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). Berlin, Germany: Springer-Verlag. pp 91102.; T. Nepusz, A. Petroczi, L. Negyessy, F. Bazso, Fuzzy communities and the concept of bridgeness in complex networks, Phys. Rev. E 77 (1) (2008) 016107; Lancichinetti A, Fortunato S, Kertesz J (2009) Detecting the overlapping and hierarchical community structure in complex networks. New J Phys 11: 033015.; vans TS, Lambiotte R (2009) Line graphs, link partitions, and overlapping communities. Phys Rev E 80: 016105. 25; Kovacs, Istvan A., et al. (2010) Community landscapes: An integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. PloS ONE 5: e12528**). 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 (**ref: Palla, Gergely, et al. "Uncovering the overlapping community structure of complex networks in nature and society." Nature 435.7043 (2005): 814-818.**). 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) (**ref: Lancichinetti, A., Radicchi, F., Ramasco, J. J., and Fortunato, S. (2011). Finding statistically significant communities in networks. PloS one, 6(4), e18961.**). 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.

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 ?

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 $t = 1\,\text{s}, 2\,\text{s}, \ldots$, 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-$k$ transfer entropy of Y on X is defined as

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

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

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

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).

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,

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

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.

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

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

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

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

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

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 (**ref: Salton, Gerard, and Michael J. McGill. "Introduction to modern information retrieval." (1986).**) 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

$$w^{\text{HT}}_{u \to f} = \frac{\vec{h}(u) \cdot \vec{h}(f)}{||\vec{h}(u)||\ ||\vec{h}(f)||}. \tag{13}$$

Results

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 , , and . The number of communities by community type is given in Table . 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 . 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.

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

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

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.**

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 2009), 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

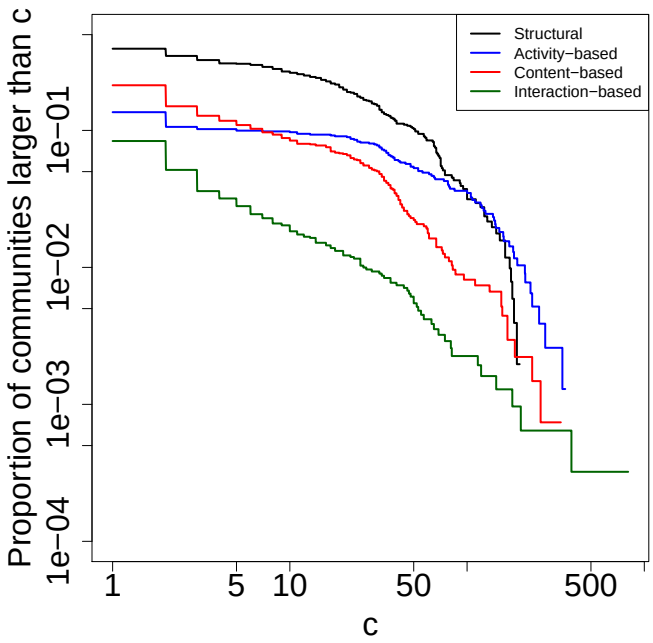

Figure 1: The proportion of communities greater than c in size, across the different community types. Note the logarithmic scale on the horizontal and vertical axes.

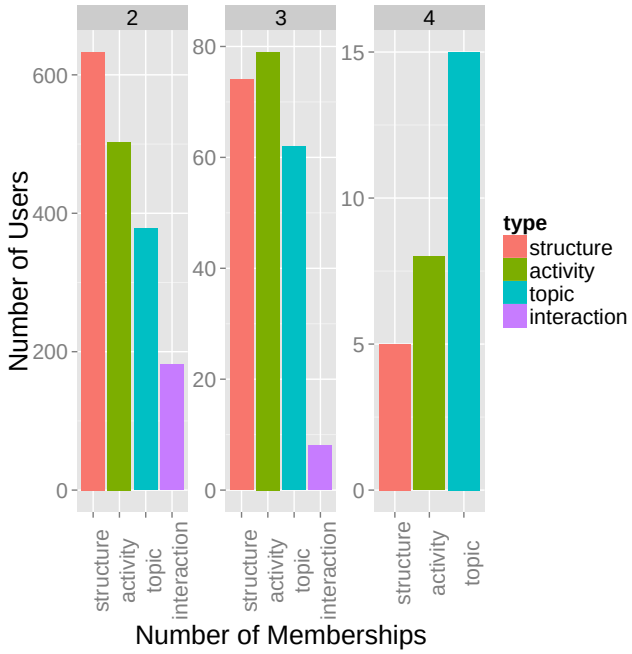

Figure 2: The number of users belonging to 2, 3, or 4 communities, by community type.

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.

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 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 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 .

We see that overall, the

Discussion

Questions Influence Community Structure

Conclusions

References

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

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

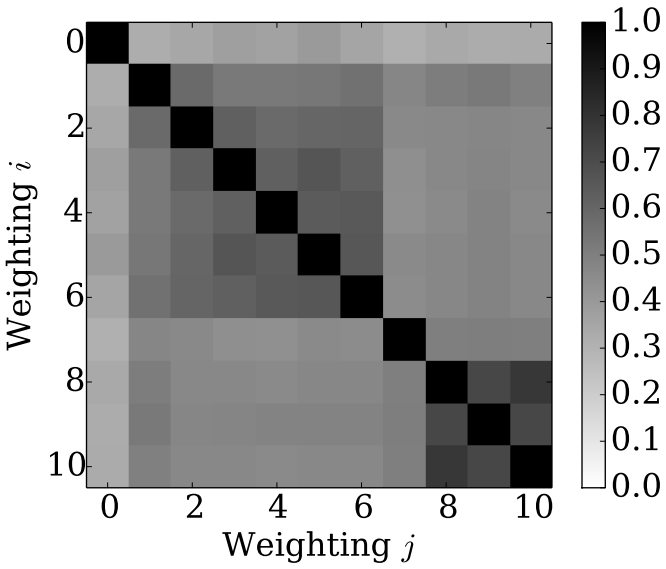

Figure 3: The normalized mutual information between the communities using the different weightings. Weighting 0 corresponds to the structural (binary weighting) network, weightings 1 through 6 correspond to the weighting using the transfer entropy with lag 1 through 6, weighting 7 corresponds to the 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.

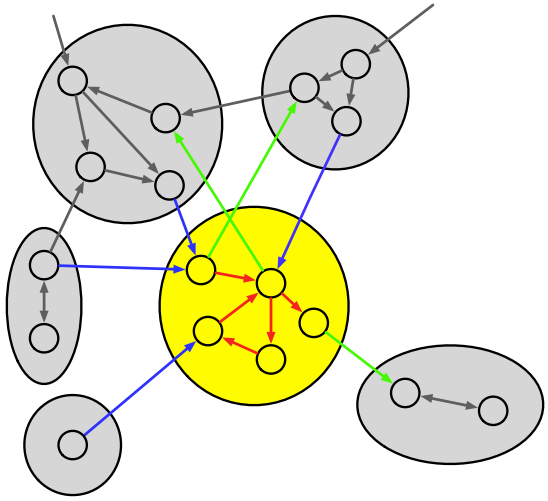

Figure 4: An example of the edges considered in determining the edge weight distribution for a given community (the focal community is in yellow). We focus on the internal-to-internal (red), internal-to-external (green), and external-to-internal (blue) edges. For a given focal community, all other edges (grey) are not considered.

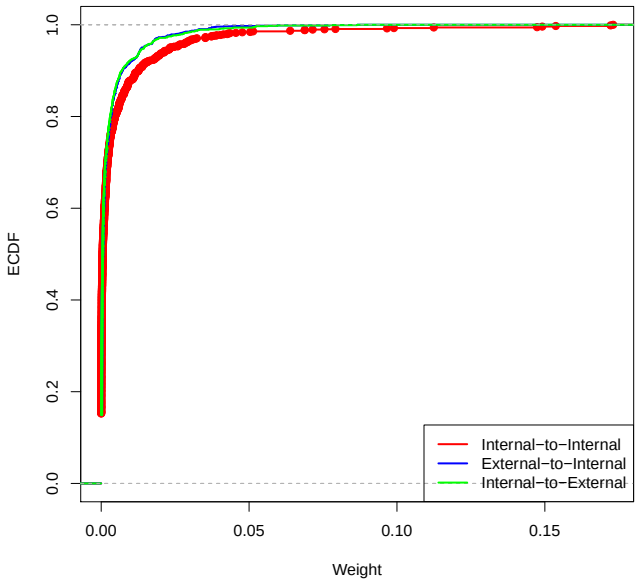

Figure 5: The proportion of edges with a weight at least as large as the weight on the horizontal axis, across the types of edges described in Figure 4. The community is defined by user interactions, and the edge weights are determined by topics.

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 networks. 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.

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

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

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

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

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

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	$0.9/0.89$	$2.4/2.6$*	$28/26$

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.

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