Community Detection Using a Measure of Global Influence

Rumi Ghosh and Kristina Lerman

USC Information Sciences Institute, Marina del Rey, California 90292
{ghosh,lerman}@isi.edu

Abstract. The growing popularity of online social networks gave researchers access to large amount of network data and renewed interest in methods for automatic community detection. Existing algorithms, including the popular modularity-optimization methods, look for regions of the network that are better connected internally, e.g., have higher than expected number of edges within them. We believe, however, that edges do not give the true measure of network connectivity. Instead, we argue that *influence*, which we define as the number of paths, of any length, that exist between two nodes, gives a better measure of network connectivity. We use the influence metric to partition a network into groups or communities by looking for regions of the network where nodes have more influence over each other than over nodes outside the community. We evaluate our approach on several networks and show that it often outperforms the edge-based modularity algorithm.

Key words: community, social networks, influence, modularity

1 Introduction

Communities and social networks have long interested researchers [5, 13]. However, one of the main problems faced by the early researchers was the difficulty of collecting empirical data from human subjects [5]. The advent of the internet and the growing popularity of online social networks changed that, giving researchers access to huge amount of social interactions data. This, coupled with the ever increasing computation speed, storage capacity and data mining capabilities, led to the reemergence of interest in the social networks in general, and community detection specifically.

Many existing community finding algorithms look for regions of the network that are better connected internally and have fewer connections to nodes outside the community [4]. Graph partitioning methods [7, 27], for example, attempt to minimize the number of edges between communities. Modularity maximization-based methods, on the other hand, identify groups of nodes that have higher than expected number of edges within them [22, 21, 24, 23]. We believe, however, that edges do not give the true measure of network connectivity. We generalize the notion of network connectivity to be the number of paths, of any length, that

exist between two nodes (Section 2). We argue that this metric, called *influence* by sociologists [13], because it measures the ability of one node to affect (e.g., send information to) another, gives a better measure of connectivity between nodes. We use the influence metric to partition a (directed or undirected) network into groups or communities by looking for regions of the network where nodes have more influence over each other than over nodes outside the community. In addition to discovering natural groups within a network, the influence metric can also help identify the most influential nodes within the network, as well as the "weak ties" who bridge different communities. We formalize our approach by describing a general mathematical framework for representing network structure (Section 3). We show that the metric used for detecting communities in random walk models, modularity-based approaches, and influence-based modularity are special cases of this general framework. We evaluate our approach (in Section 4) on the standard data sets used in literature, and find performance at least as good as that of the edge-based modularity algorithm.

2 A Measure of Global Influence

If a network has N nodes and E links it can be graphically represented by $G(N, E)$ where N is the number of vertices and E is the number of edges. Edges are directed; however, if there exists an edge from vertex i to j and also from j to i, it is represented as an undirected edge. A path p is an *n-hop path* from i to j, if there are n vertices between the i and j along the path. We allow paths to be non-selfavoiding, meaning that the same edge could be traversed more than once. The graph $G(N, E)$ can be represented by an adjacency matrix A, whose elements are defined as $A_{ij} = 1$ if $\exists$ an edge from vertex i to j; otherwise, $A_{ij} = 0$. A is symmetric for undirected graphs.

The *Oxford English Dictionary* defines influence as "the capacity to have an effect on the character, development, or behavior of someone or something, *or* the effect itself." The measure of influence that we adopt is along lines of Pool and Kochen [5], who state that "influence in large part is the ability to reach a crucial man through the right channels, and the more the channels in reserve the better." This metric depends not only on direct edges between nodes, but also on the number of ways an effect or a message can be transmitted through other nodes. Therefore, the capacity of node i to influence node j can be measured by the weighted sum of the number of *n-hop* paths present from the i to j. The underlying hypothesis is that the more the number of paths from one node to another, the greater is the capacity to influence. This model is analogous to the independent cascade model of information spread [11, 14].

The strength of the effect via longer paths is likely to be lower than via shorter paths. We model the attenuation using parameters α_i where α_i $(0 \leq \alpha_i \leq 1)$ is the probability of transmission of effect in the (i-1)th hop. Let us consider transmitting an effect or a message from nodes b to c in the network shown in Figure 1. The probability of transmission via the immediate neighbors of c such as e to c or g to c is α_1. The probability of transmission over 1-hop paths such

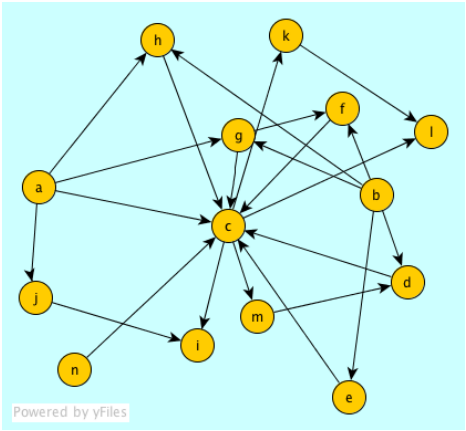


Fig. 1. A directed graph representing a network

as b to c via e is $\alpha_1\alpha_2$. In general, the probability of a transmission along an n-hop path is $\Pi_{i=1}^{n+1}\alpha_i$. The total influence of b on c thus depends on the number of (attenuated) channels between b and c, or the sum of all the weighted paths from b to c. This definition of influence makes intuitive sense, because the greater the number of paths between b and c, the more opportunities there are for b to transmit messages to c or to affect c.

For ease computation we simplify this model by taking $\alpha_1 = \beta$ and $\alpha_i = \alpha$, $\forall i \neq 1$. β is called the direct attenuation factor and is the probability of transmission of effect directly between adjacent nodes. α is the indirect attenuation factor and is the probability of transmission of effect through intermediaries. If $\alpha = \beta$, i.e., the probability of transmission of effect through all links is the same, then this index reduces to the metric used to find the Katz status score [13].

The number of paths from i to j with n intermediaries, $i \xrightarrow{n} j$, is given by $A_n = \overbrace{A \cdot A \cdots A}^{n+1 \ times} = A_{(n-1)} \cdot A$. Adding weights to take into account the attenuation of effect, we get the weighted total capacity of i to affect j as $i \longrightarrow j \ = \beta \ i \xrightarrow{0} j \ + \cdots + \beta\alpha^n \ i \xrightarrow{n} j \ + \cdots$. We represent this weighted total capacity to influence by the *influence matrix* P:

$$\begin{aligned} P &= (\beta A + \beta\alpha A_1 + \cdots + \beta\alpha^n A_n + \cdots) \\ &= \beta A(I - \alpha A)^{-1}, \end{aligned} \tag{1}$$

As mentioned by Katz[13], the equation holds while $\alpha < 1/\lambda$, where λ is the largest characteristic root of A [6].

We use the influence matrix to help find community structure in a network. We claim (without much theoretical or empirical support) that a *community is composed of individuals who have a greater capacity to influence others within their community than outsiders.* As a result, actions of community members will

tend to become correlated with time, whether by adopting a new fashion trend, vocabulary, watching a movie, or buying a product. Armed with this alternative definition of community, we adapt modularity maximization-based approach to identifying communities.

2.1 Influence-based Modularity

The objective of the algorithms proposed by Newman and coauthors is to discover "community structure in networks — natural divisions of network nodes into densely *connected* subgroups" [25]. They proposed *modularity* as a measure for evaluating the strength of the discovered community structure. Algorithmically, their approach is based on finding groups with higher than expected edges within them and lower than expected edges between them [22, 21, 23]. The modularity Q optimized by the algorithm is given by:
Q =(fraction of edges within community)-(expected fraction of such edges). Thus, Q is used as a numerical index to evaluate a division of the network. The underlying idea, therefore, is that connectivity of nodes within a community is greater than that of nodes belonging to different communities, and they take the number of edges as the measure of connectivity. However, we claim that path-based, rather than edge-based, connectivity is the true measure of network connectivity. Consider again the graph in Figure 1, where there exists an edge from a to c but not from b to c. Clearly, however, c is not unconnected from b, as there are several distinct paths from b to c. We use the influence matrix, which gives the global connectivity of the network, to identify communities.

We redefine modularity as Q =(connectivity within the community) - (expected connectivity within the community) and adopt the influence matrix P as the measure of connectivity. This definition implies that in the best division of the network, the influence of nodes within their community is greater than their influence outside their community. A division of the network into communities, therefore, maximizes the difference between the actual capacity to influence and the expected capacity to influence, given by the capacity to influence in an equivalent random graph.

Let us denote the expected capacity to influence by an $N \times N$ matrix $\bar{P}$. We round off the values of P_{ij} to the nearest integer values. Modularity Q can then be expressed as

$$Q = \sum_{ij} [P_{ij} - \bar{P_{ij}}]\delta(s_i, s_j) \tag{2}$$

where s_i is the index of the community i belongs to and $\delta(s_i, s_j) = 1$ if $s_i = s_j$; otherwise, $\delta(s_i, s_j) = 0$. When all the vertices are placed in a single group, then axiomatically, $Q = 0$. Therefore $\sum_{ij}[P_{ij} - \bar{P_{ij}}] = 0$. Hence, the total capacity to influence W is

$$W = \sum_{ij} \bar{P_{ij}} = \sum_{ij} P_{ij} \tag{3}$$

Hence the null model has the same number of vertices N as the original model, and in it the expected influence of the entire network equals to the actual influence of the original network. We further restrict the choice of null model to that

where the expected influence on a vertex j, W_j^{in}, is equal to the actual influence on the corresponding vertex in the real network.

$$W_j^{in} = \sum_i \bar{P_{ij}} = \sum_i P_{ij} \tag{4}$$

Similarly, we also assume that in the null model, the expected capacity of a vertex i to influence others, W_i^{out}, is equal to the actual capacity to influence of the corresponding vertex in the real network

$$W_i^{out} = \sum_j \bar{P_{ij}} = \sum_j P_{ij}. \tag{5}$$

In order to compute the expected influence, we reduce the original graph G to a new graph G' that has the same number of nodes as G and total number of edges W, such that each edge has weight 1 and the number of edges between nodes i and j in G' is P_{ij}. So now the expected influence between nodes i and j in graph G could be taken as the expected number of the edges between node i and j in graph G' and the actual influence between nodes i and j in graph G can be taken as the actual number of edges between nodes i and node j in graph G'. The equivalent random graph G'' is used to find the *expected* number of edges from node i to node j. In this graph the edges are placed at random subject to constraints:

- The total number of edges in G'' is W.
- The out-degree of a node i in G'' = out-degree of node i in $G' = W_i^{out}$.
- The in-degree of a node j in graph G'' =in-degree of node j in graph $G' = W_j^{in}$.

Thus in G'' the probability that an edge will emanate to a particular vertex i is dependent only on the out-degree of that vertex; and the probability that an edge is incident on a particular vertex i is dependent only on the in-degree of that vertex and the probabilities of the two vertices being the two ends of a single edge are independent of each other. In this case, the probability that an edge exists from i to j is given by $P(\textit{emanates from } i) \cdot P(\textit{incident on } j)=(W_i^{out}/W)(W_j^{in}/W)$. Since the total number of edges is W in G'', therefore the expected number of edges between i and j is $W \cdot (W_i^{out}/W)(W_j^{in}/W) = \bar{P_{ij}}$, the expected influence between i and j in G.

2.2 Detecting Community Structure

Once we have derived Q, we have to select an algorithm to divide the network into communities that optimize Q. Brandes et al. [3] have shown that the decision version of modularity maximization is NP-complete. Like others [23, 17], we use the the leading eigenvector method to obtain the approximate solution. In [9] we applied this approach to the standard data sets used in literature, and found performance at least as good as that of the edge-based modularity algorithm. As can be mathematically derived from the formulation, we find that the

communities detected are independent of the value of β. So, henceforth without loss of generality, we shall assume the value of $\beta = 1$. In Section 4 we use this approach to partition several example networks into communities.

3 A Generalized Model of Influence

In this section, we present a mathematical framework that generalizes the notion of influence. In algebraic topology a *k-simplex*, with $k \geq 0$, is a convex hull σ of $k+1$ linearly independent points $v_0, v_1, \ldots, v_k$ and dimension k. The points v_i are called vertices of σ. Let $\sigma = \{v_0, v_1, \ldots, v_k\}$ be a k-simplex and let $\omega = \{w_i, \ldots, w_l\}$ be a nonempty subset of σ, where $w_i \neq w_j$ if $i \neq j$. Then $\omega = \{w_0, w_1, \ldots, w_l\}$ is called the l-dimensional *face* of σ. A *simplectic complex* K is a finite collection of simplices in some R^n satisfying:

- If $\sigma \in K$, then all faces of σ belong to K.
- If $\sigma_1, \sigma_2 \in K$, then either $\sigma_1 \bigcap \sigma_2 = \emptyset$ or $\sigma_1 \bigcap \sigma_2$ is a common face of σ_1 and σ_2 .

The dimension of K is defined to be -1 if $K = \emptyset$ and the maximum of the dimensions of K otherwise. An undirected graph can then be viewed as a simplectic complex with a single-element set per vertex and a two-element set per edge.

Suppose we are given finite sets $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_m\}$, and a binary relation $\gamma \subseteq X \times Y$ between elements of X and elements of Y (X and Y could be the same). Then the relation γ may be expressed as an $n \times m$ incidence matrix $A'(\gamma) = (A'_{ij})$ where $A'_{ij} = 1$ if $(x_i, y_j) \in \gamma$ and 0 otherwise. Each row in the incidence matrix $A'(\gamma)$ may be viewed as a simplex in the following way: Let Y be a set of vertices. The i-th row of $A'(\gamma)$ can be identified with a k dimensional simplex $\{y_{j1}, y_{j2}, \ldots, y_{j(k+1)}\} = \sigma_k(x_i)$ on the vertices Y (where A'_{ij} =1). Thus each $x_i \in X$ determines (with γ) a row of $A'(\gamma)$ and each row $A'(\gamma)$ can be identified a simplex. The set of simplices is a simplical complex denoted by $K_X(\gamma, Y)$. Since an arbitrary element x_i is γ-related to exactly $k+1$ y_j, $\sigma_k(x_i)$ is distinguished as a named simplex. If we let d denote the maximum dimension of $K_X(\gamma, Y)$, we immediately see that $d \leq m-1$.

Let σ and τ be two simplices in $K_X(\gamma, Y)$. Then σ and τ are $q-near$ if they have a common q-face, i.e., their intersection contains at least $q+1$ elements. (This q-face need not be an element of the simplex family.) Then τ and σ are q-connected if there exists a sequence $\sigma_1, \sigma_2, \ldots, \sigma_p$ of simplices in $K_X(\gamma, Y)$, such that $\sigma_1 = \sigma$ and $\sigma_p = \tau$ and σ_i is q-near to σ_{i+1} $\forall 1 \leq i \leq p-1$. Thus q-connectivity is the transitive closure of q-nearness. Q analysis using q-nearness and q-connectivity was used by Atkin [2] to deal with pairs of sets and sets of contextual relations.

As Legrand [16] points out, q-nearness and q-connectivity are not necessarily a true measure of how similar the vertices are to each other, for which the length of sequence of q-connectivity should be the true indicator. We therefore take the length of sequences into account by calculating how q-near vertex i is

from vertex j, making it dependent on the length of the path between them. Therefore the adjacency matrix A, $A_{ij} = (q1_{ij})$, shows if two simplices are zero-near to one another in a 0-hop path. The product $A^2 = A \times A$ gives the value of $q2_{ij}$ such that ${A^2}_{ij} = q2_{ij}$, i.e., vertex i and vertex j when separated by a one-hop path are q-near each other with $q = q2_{ij} - 1$. In the same way, $A^3 = A \times A \times A = (A^3_{ij}) = (q3_{ij})$ shows that vertices i and j connected by a two-hop path are $q3_{ij} - 1$ near from each other. We then take the length of the sequence into account to calculate the expected q-nearness of one vertex to another by taking the weighted average of q-nearness of varying length of paths. The expected value of q_{ij} between two elements i and j, such that they are expected to be $q_{ij} - 1$ near each other, with $(qk_{ij}) = A^k_{ij}$ is:

$$\mathbf{E(q_{ij})} = \frac{(\mathbf{W_1 \cdot q1_{ij} + W_2 \cdot q2_{ij} + \ldots + W_n \cdot qn_{ij} + \ldots})}{\sum_{\mathbf{i=1}}^{\infty} \mathbf{W_i}} \tag{6}$$

This expected value can be used to find out how connected two vertices are to each other, taking paths of all lengths into account. Note that W_i can be a scalar or a vector.

This formulations allows us to generalize different network models for community detection and scoring like the random walk model [28, 29, 31], the Katz model [13] of status score, and the influence-based model. In random walk models, a particle starts a random walk from node i. The particle iteratively transitions to its neighbors with probability proportional to the corresponding edge weights. Also at each step, the particle returns to node i with some restart probability $(1 - c)$. The proximity score from node i to node j is defined as the steady-state probability $r_{i,j}$ that the particle will be on node j [29]. These models can be shown to be special cases of the formulations of the expected q-nearness (without loss of generality we assume that T is an $n \times n$ matrix):

1. If $W_k = c^{k-1} \cdot D^{-(k-1)}$ where c is a constant and D is an $n \times n$ matrix with $D_{ij} = \sum_{j=1}^{n} A_{ij}$ if $i = j$ and 0 otherwise; then, the expected q-nearness score reduces to proximity score in random walk model [28, 29].
2. If $W_i = \Pi_{j=1}^{i} \alpha_j$, where the scalar α_j is the attenuation factor of a $(j-1)$-th hop in a $(i-1)$ hop path, then the expected q-nearness reduces to metric used to find the influence score and represented by the influence matrix. For ease of computation of the influence matrix, we have taken $\alpha_1 = \beta$ and $\alpha_i = \alpha$ $\forall i \neq 1$. As stated before, $\alpha < 1/\lambda$ where λ is the largest characteristic root of A. Gershgorins Circle Theorem (1931) gives the simple sufficient condition $\alpha < 1/max_i(D_{ii})$.
3. When $\beta = \alpha$, this in turn reduces to the metric used to find the Katz status score [13] with α as the attenuation factor.
4. When $\alpha_1 = 1$ and $\alpha_2 = \ldots = \alpha_n = \ldots = 0$, the expected q-nearness is the q-nearness of the 0-hop path which is metric used to calculate similarity in edge-based modularity approaches [22].

In summary, the capacity to influence is a measure of the expected q-nearness between vertices. Liben-Nowell and Kleinberg [20] have shown that Katz measure

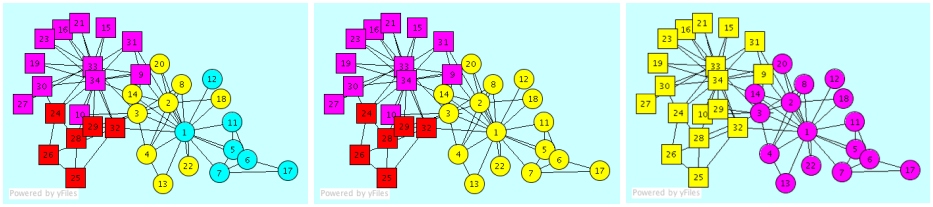
is the most effective measure for link prediction task. The influence score, which is a generalization of the Katz score, can then be used to find communities as described in Section 2.1.

4 Evaluation

We applied influence-based community finding method to small networks studied previously in literature, as well as the friendship network extracted from the social photosharing site Flickr. On all the data sets we studied, the performance of the influence-based modularity optimization algorithm was at least as good as that of the edge-based modularity ($\alpha = 0$ case). In several cases, the influence-based approach led to purer groups.

4.1 Zachary's Karate Club

The karate club data represents the friendship network of members of a karate club studied by Zachary [30]. During the course of the study, a disagreement developed between the administrator and the club's instructor, resulting in the division of the club into two factions, represented by circles and squares in Figure 2. We used this data to study the communities detected for different values of α and compare the performance of the influence-based modularity approach to Newman's community-finding algorithms (which is the special case where $\alpha = 0$) [21].



(a) (b) (c)

Fig. 2. Zachary's karate club data. Circles and squares represent the two actual factions, while colors stand for discovered communities as the strength of ties increases: (a)$\alpha = 0$ (b) $0 < \alpha < 0.14$ (c)$0.14 \leq \alpha \leq 0.29$

Using $\alpha < 1/\lambda$ (Section 2) we get the upper bound on $\alpha \leq 0.29$. When both Newman's edge-based modularity maximization approach and our method ($0 \leq \alpha \leq 0.29$) are used to bisect the network into just two communities, we recover the two factions observed by Zachary (Figure 2(c)). However, algorithms run until a termination condition is reached (no more bisections are possible), different values of α lead to different results, as shown in Figure 2. As stated when $\alpha = 0$, the method reduces to Newman's edge-based modularity maximization

approach [21], and we get four communities (Figure 2(a)). For $0 < \alpha < 0.14$ the number of communities reduces to three (Figure 2(b)). As α is increased further ($0.14 \leq \alpha \leq 0.29$) we get two communities(Figure 2(c)), which are the same as the factions found in Zachary's study.

4.2 College Football

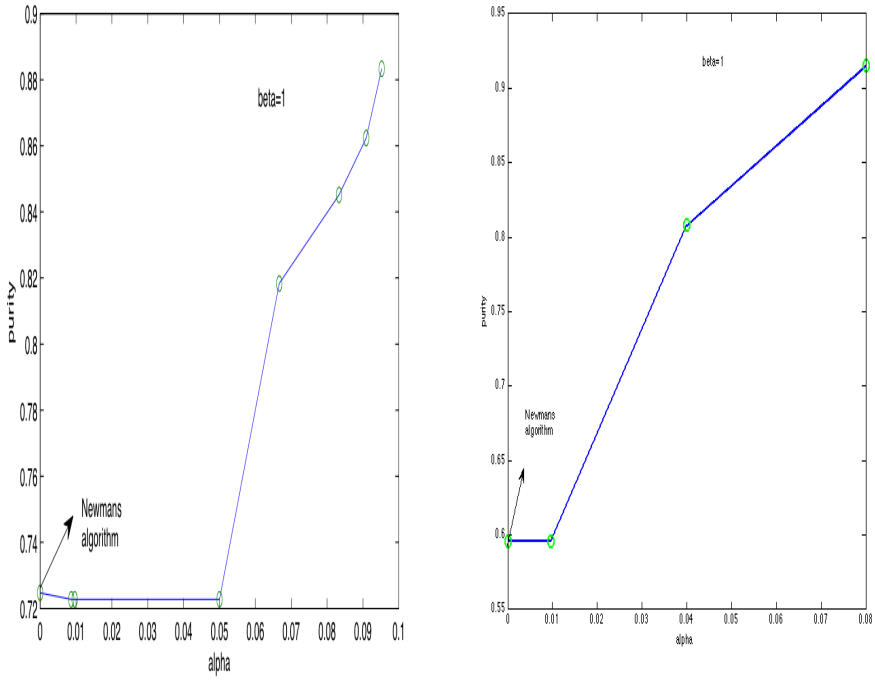
We also ran our approach on the US College football data from Girvan et al. [10].[1] The network represents the schedule of Division 1 games for the 2000 season where the vertices represent teams and the edges represent the regular season game between the two teams they connect. The teams are divided into "conferences" (or communities) containing 8 to 12 teams each. Games are more frequent between members of the same conference than members of different conferences. Inter-conference games, however, are not uniformly distributed, with teams that are geographically closer likely to play more games with one another than teams separated by geographic distances. However, some conferences have teams playing nearly as many games against teams in other conferences as teams within their own conference. This leads to the intuition, that conferences may not be the natural communities, but the natural communities may actually be bigger in size than conferences, with teams playing as many games against others in the same conferences being put into the same community.

We measure the quality of discovered communities in terms of purity. The purity of a community is the fraction of all pairs of teams that were assigned to that community that actually belong to the same conference. The quality of a network division produced by an algorithm is the average purity of the discovered communities. Figure 3 shows the purity the discovered communities as α is varied. Purity is independent of β, the weight of direct edges, but increases with α, reaching $\sim 88\%$ near $\alpha = 0.1$ (the upper bound to α is determined by the reciprocal of the largest eigenvalue of the adjacency matrix). When $\alpha = 0$, the modularity reduces to edge-based modularity studied by Newman [23], the purity is around 72%. The number of predicted groups changes from 8 at $\alpha = 0$ to four at higher values of α.

4.3 Political Books

We evaluated our approach on the political books data compiled by V. Krebs.[2] In this network the nodes represent books about US politics sold by the online bookseller Amazon. Edges represent frequent co-purchasing of books by the same buyers, as indicated by the "customers who bought this book also bought these other books" feature of Amazon. This feature influences the book purchasing decisions of customers. The nodes were given labels *liberal*, *neutral*, or *conservative* by Mark Newman on a reading of the descriptions and reviews of

[1] `http://www-personal.umich.edu/~mejn/netdata/`
[2] `http://www.orgnet.com/`



College football Political books

Fig. 3. The graph showing the purity of communities predicted with different values of α ($\beta = 1$)in the (a) college football and (b) political books data sets. We see that purity increases with α. When $\alpha = 0$, the method reduces to eigenvector based modularity maximization method postulated by Newman [23].

the books posted on Amazon.[3] 49 of the books were marked as *conservative*, 43 books were marked as *liberal* and 13 books were marked as *neutral*. We use our algorithm to find the existing community structure in the network by varying parameters as shown in Figure 3. Purity is independent of the value of β (and hence β is taken as 1), and similarly to the football data, as α increases, the number of communities decreases (from four at $\alpha = 0$ to two at $\alpha = 0.08$). Also the purity of the communities increases from 60% at $\alpha = 0$ to 92% at $\alpha = 0.08$. Again, $\alpha = 0$ corresponds to Newman's modularity method. Another observation is that when $\alpha = 0.08$, leading to the formation of two groups, only the neutral books are split between group, indicates a possibility that some of the 13 *neutral* books were conservatively inclined and some liberally.

4.4 Flickr Social Network

We also ran our algorithm on the social network data collected from Flickr for the image search personalization study [18]. Flickr is a social photosharing site that allows users to upload images, tag them with descriptive keywords, known as tags, and to join social networks by adding other users as contacts. We believe that network structure, create by independent decisions to add another photographer as a contact, capture social knowledge, including knowledge about users' photography interests. Thus, users who are interested in a particular topic are more likely to be connected than users interested in different topics.

Since the actual social network on Flickr is rather vast, we sampled it by identifying users who were broadly interested in one of three topics: child and family *portraiture*, *wildlife* photography and *technology*. For each topic, we used the Flckr API to perform a tag search using a keyword relevant to that topic, and retrieved 500 'most interesting' images. We then extracted the names of users who submitted these images to Flickr. These users were added to our data set. The keywords used for image search were (a) newborn for the *portraiture* topic, (b) tiger and beetle for the *wildlife* topic, and (c) apple for the *technology* topic. Each keyword is ambiguous. Tiger, for example, could mean a big cat of the *panthera* genus, but also a flower (Tiger lily), Mac operating system (OS X Tiger), or a famous golfer (Tiger Woods), while beetle could describe a bug or a car. The keyword newborn could refer to human babies just as well as to kittens and puppies, while apple could mean the computer maker or a fruit.

From the set of users in each topic, we identified four (eight for the *wildlife* topic) who were interested in the topics we identified: i.e., *wildlife* for tiger and beetle query terms, *portraiture* for the newborn query, and *technology* for the apple query. We studied each user's profile to confirm that the user was indeed interested in that topic. Specifically, we looked at group membership and user's most common tags. Thus, groups such as "Big Cats", "Zoo", "The Wildlife Photography", etc. pointed to user's interest in the *wildlife* topic. In addition to group membership, tags that users attached to their images could also help identify their interests. For example, users who used tags nature and macro were

[3] available at `http://www-personal.umich.edu/~mejn/netdata/`

probably interested *wildlife* rather than *technology*. Similarly, users interested in human, rather than animal, *portraiture* tagged their images with baby and family. We used the Flickr API to retrieve the contacts of each of the users we identified, as well as their contacts' contacts. We labeled users by the topic through which they were discovered. In other words, users who uploaded one of the 500 most interesting images retrieved by the query tiger, were labeled *wildlife*, whether or not they were interested in wildlife photography. The contacts and contacts's contacts of the four users within this set identified as being interested in wildlife photography were also labeled *wildlife*. Although we did not verify that all the labeled users were indeed interested in the topic, we use these soft labels to evaluate the discovered communities.

Once we retrieved the social networks of target set of users, we reduced it to an undirected network containing mutual contacts only. In other words, every link in the network between two nodes, say A and B, implies that A lists B as contact and *vice versa*. This resulted in a network of 5747 users. Of these, 1620 users were labeled *technology*, 1337 and 2790 users were labeled *portraiture* and *wildlife* respectively. We ran our community finding algorithm for different values of α on this data set. For $\alpha = 0$, we found four groups, while for higher values of α ($\alpha < 0.01$), we found three groups. Figure 4 shows composition of the discovered groups in terms of soft labels. Group 1 is composed mainly of *technology* users, group 2 mainly *wildlife* users, and group 3 is almost exclusively *portraiture*. The fourth group found at $\alpha = 0.0$ has 932 members, of which 497 are labeled *wildlife*, 242 *technology*, and 193 members *portraiture*. Except for the *portraiture* group (group 3), groups become purer as α increases.

5 Related Research

There has been some work in motif-based communities in complex networks [1] which like our work extends traditional notion of modularity introduced by Girvan and Newman [10]. The underlying motivation for motif-based community detection is that "the high density of edges within a community determines correlations between nodes going beyond nearest-neighbours," which is also our motivation for applying the influence-based modularity metric to community detection. Though the motivation of this method is to determine the correlations between nodes beyond nearest neighbors, yet it does impose a limit on the proximity of neighbors to be taken into consideration dependent on the size of the motifs. The method we propose, on the other hand, imposes no such limit on proximity. On the contrary, it considers the correlation between nodes in a more global sense. The measure of global correlation evaluated using the influence metric would be equal to the weighted average of correlations when motifs of different sizes are taken. The influence matrix enables the calculation of this complex term in a quick and efficient manner.

Resolution limit is one of the main limitations of the original modularity detection approach [8]. It can account for the comment by Leskovec et al. [19] that they "observe tight but almost trivial communities at very small scales, the

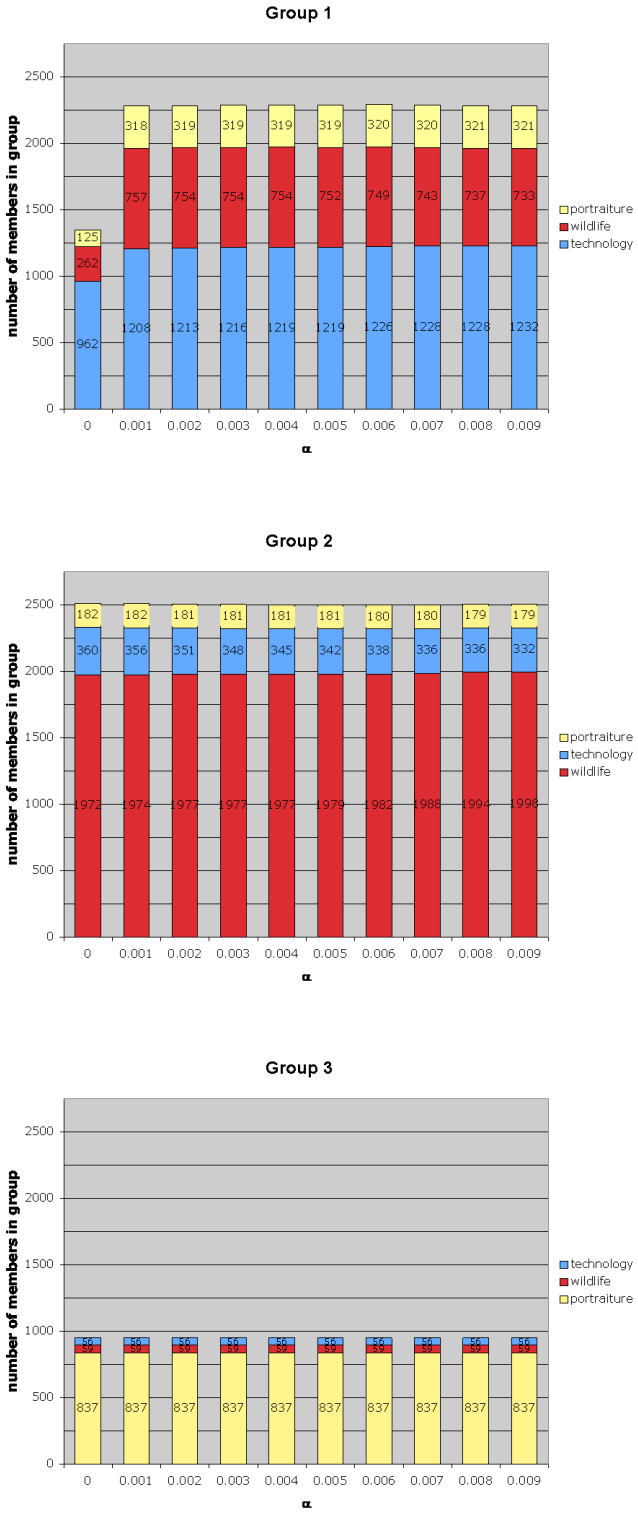


Fig. 4. Composition of groups discovered in the Flickr social network for different values of α

best possible communities gradually 'blend in' with rest of the network and thus become less 'community-like'." However, that study is based on the hypothesis that communities have "more and/or better-connected 'internal edges' connecting members of the set than 'cut edges' connecting to the rest of the world." Hence, like most graph partitioning and modularity-based approaches to community detection, their process depends on the local property of connectivity of nodes to neighbors via edges and is not dependent on the structure of the network on the whole. Therefore, it does not take into account the characteristics of node types, that is 'who' are the nodes that a node is connected to and how influential these nodes are. In their paper on motif-based community detection, Arenas et al.[1] state that the extended quality functions for-motif based modularity also obey the principle of the resolution limit. But this limit is now motif-dependent and then several resolution of substructures can be achieved by changing the motif. However, it would be difficult to verify which resolution of substructures is closest to natural communities. In influence-based modularity, on the other hand, the resolution limit would depend on the probability of transmission of the effect between nodes, i.e., the strength of ties. The probability of transmission of effect can indeed be calculated from the graph, by say observing the dynamics of spread of idea within a graph at different times.

As stated before, Liben-Nowell and Kleinberg [20] have shown that Katz measure is the most effective measure for the link prediction task, better than hitting time, PageRank [26] and its variants. Thus we use influence score, which is a generalization of the Katz score, to detect communities and compute rankings of individuals.

Recently researchers have used probabilistic models, e.g., mixture models, for community discovery. These models can probabilistically assign a node to more than one community, as it has been observed "objects can exhibit several distinct identities in their relational patterns" [15]. This indeed may be true, but whether the nodes in the network is to be divided into distinct communities or probabilities with which each node belongs to community is to be discovered, really depends on the specific application. By this, we mean that if the application we are interested in is finding the natural communities say in the karate club data, and if we use a probabilistic method (say [15]), we would be assigning the nodes into groups into which their probability of belonging is the highest, and the communities thus formed do not necessarily portray the division of the network into natural communities observed.

6 Conclusion and Future Work

We have proposed a new definition of community in terms of the capacity of nodes to influence each other. We gave a mathematical formulation of this effect in terms of the number of paths of any length that link two nodes, and redefined modularity in terms of the influence metric. We use the new definition of modularity to partition a network into communities. We applied this framework to networks well-studied in literature and found that it produces results at least as

good as the edge-based modularity approach. We were able to apply this framework to data extracted from the online social networking site Flickr to get very promising results.

Although the formulation developed in this paper applies equally well to directed graphs, we have only implemented it on undirected ones. Hence future work includes implementation of the of the algorithm on directed graphs that are common on social networking sites, as well applying it to bigger networks. The influence matrix approximates capacity to influence along the lines of independent cascade model of information spread. Future work includes approximation of capacity to influence along other models of information spread like the threshold influence model.

Also, in the lines of the Katz score, influence metric allows us to compute the rank of nodes relative to each other, these rankings will depend on the value of α, the indirect attenuation factor. Preliminary investigation of these influence based ranking scores has been very encouraging. Not only has it enabled us to identify the most influential nodes, but has also shown that as the weight of indirect links grows, the rank of the nodes that act as bridges between communities increase. The influence metric thus allows us to identify who the "weak ties" [12] are along with the influential ties. Future direction includes the study of influence based ranking and the interesting network structural properties it reveals.

Acknowledgements

This research is based on work supported in part by the National Science Foundation under Award Nos. IIS-0535182, BCS-0527725 and IIS-0413321.

References

1. Alex Arenas, Alberto Fernandez, Santo Fortunato, and Sergio Gomez. Motif-based communities in complex networks. *Mathematical Systems Theory*, 41:224001, 2008.
2. R Atkin. From cohomology in physics to q-connectivity in social science. *International Journal of Man-Machines Studies*, 4:341–362, 1972.
3. U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *IEEE Trans. on Knowl. and Data Eng.*, 20(2):172–188, 2008.
4. A. Clauset. Finding local community structure in networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 72(2), 2005.
5. I. de Sola Pool and M. Kochen. Contacts and influence. *Social Networks*, 1(1):39–40, 1978–1979.
6. W. L. Ferrar. *Finite Matrices*. Oxford Univ. Press, 1951.
7. M. Fiedler. Algebraic connectivity of graphs. *Czech. Math. J.*, 23:298–305, 1973.
8. Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *PROC.NATL.ACAD.SCI.USA*, 104:36, 2007.
9. R. Ghosh and K. Lerman. Community detection using a measure of global influence. In *Proc. of the 2nd KDD Workshop on Social Network Analysis (SNAKDD'08)*, 2008.

10. M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PROC. NATL. ACAD. SCI. USA*, 99:7821, 2002.
11. Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 2001.
12. M. Granovetter. The strength of weak ties. *The American Journal of Sociology*, May 1973.
13. L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–40, 1953.
14. David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, New York, NY, USA, 2003. ACM.
15. P.S. Koutsourelakis and T. Eliassi-Rad. Finding mixed-memberships in social networks. *AAAI Spring Symposium Social Information Processing*, 2008.
16. J Legrand. How far can q-analysis go into social systems understanding? *Fifth European Systems Science Congress*, 2002.
17. E. A. Leicht and M. E. J. Newman. Community structure in directed networks. *Physical Review Letters*, 100:118703, 2008.
18. K. Lerman, A. Plangprasopchok, and C. Wong. Personalizing results of image search on flickr. In *AAAI workshop on Intelligent Techniques for Web Personlization*, 2007.
19. J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceedings of the World Wide Web Conference*, 2008.
20. David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, 58(7):1019–1031, 2007.
21. M. E. J. Newman. Detecting community structure in networks. *The European Physical Journal B*, 38:321–330, 2004.
22. M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69:066133, 2004.
23. M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74:036104, 2006.
24. M. E. J. Newman. Modularity and community structure in networks. *PROC. NATL. ACAD. SCI. USA*, 103:8577, 2006.
25. M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.
26. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
27. A. Pothen, H. Simon, and K.P. Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal. Appl.*, 11:430–452, 1990.
28. H. Tong, C. Faloutsos, and J. Pan. Fast random walk with restart and its applications. *Data Mining, 2006. ICDM '06. Sixth International Conference on*, pages 613–622, Dec. 2006.
29. Hanghang Tong, Spiros Papadimitriou, Philip S. Yu, and Christos Faloutsos. Proximity tracking on time-evolving bipartite graphs. In *SDM*, pages 704–715. SIAM, 2008.
30. W. W. Zachary. An information ow model for conict and ssion in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.
31. H Zhou. Network landscape from a brownian particles perspective. *Physical Review E*, 67, 2003.