Friends, Enemies, and Lovers: Detecting Communities in Networks Where Relationships Matter

David Jurgens[1,2]
[1]Department of Computer Science
University of California, Los Angeles
jurgens@cs.ucla.edu

Tsai-Ching Lu[2]
[2]HRL Laboratories, LLC
Malibu, California, USA
tlu@hrl.com

WebSci 2011, June 22–24, 2012, Evanston, Illinois, USA.
ACM 978-1-4503-0268-5/11/05.

Abstract

Social relationships matter: being connected with another person as a friend, enemy, or lover conveys very different information. However, current community detection methods simplify these relationships into binary connections, thereby ignoring important distinctions in how entities are connected. We highlight a new challenge for community detection on multiplex networks where entities share one or more edges, indicating different relationships. Further, we propose a new algorithm for finding communities in such networks and show promising performance on synthetic and real-world networks.

Author Keywords

community detection; social relationships; mulitplex networks

ACM Classification Keywords

H.5.2 [Information Interfaces and Presentation: User interfaces]: Evaluation/ methodology.

General Terms

Algorithms; Design

Introduction

Networks are an essential representation for modeling entities and how those entities are connected.

Furthermore, the network's edges may be annotated by edge colors or types that indicate specific types of connections between entities, e.g., people connected by a different social relationships, organisms by different environmental roles, or countries by different alliances and treaties. These different types of connections carry with them important distinctions, both in how the entities relate with one another but also in how that connection fits in the global organization of relationships.

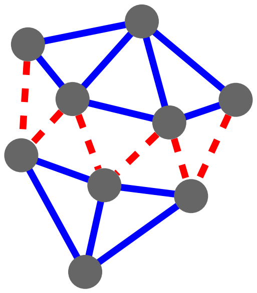

Figure 1: A hypothetical social network modeling friends and enemies. Solid lines (blue) indicate friends, while dashed lines (red) indicate enemies.

Community detection aims to use these network connections in order to identify groups of vertices that correspond to meaningful subcomponents of the network, e.g., groups of friends in a social network. Traditionally, community detection methods have focused on networks where nodes are connected by a single edge, which is possibly weighted (c.f. [5, 7] for recent surveys of community detection methods). However, many types of real-world networks are incorrectly modeled as unimodal networks with only one type of edge. For example, Figure 1 illustrates a simple, hypothetical social network with two groups of friends who are enemies of each other. Whereas the edge types reveal a clear separation in the two groups, any approach that treats edges uniformly would have significant difficulty separating the two groups solely on the basis of their network connectivity.

While network modelers have long used edge types to convey important relational distinctions, only recently have community detection methods begun to leverage the information in edge types [6, 8, 11, 10, 2]. Explicitly modeling edge types offer two important benefits. First, because the types can convey information about the nature of the interactions, community detection can better partition the network in such a way that reflects the coherency of the communities' connections. Second, edge types provide a natural scaffolding to perform community detection on multigraphs, or multiplex networks, where entities may be connected by one or more types of relations. These multiplex edges also reveal communities that invisible to a single-mode analysis, such embedded or strongly overlapping communities connected by different edge type than the rest of the network, or communities where connections between entities follow patterns in which relations they share.

Therefore, we propose a new method for community detection that uses a multi-modal *relation*-based analysis instead of the common binary connectivity analysis. We then analyze a synthetic and real-world network using the algorithm, illustrating where edge type information reveals significantly different and more informative communities than if type information was not used. Last, we outline several avenues for future research.

Relation-based Community Detection

We propose a novel approach to community detection in multiplex networks that uses both local connectivity as well as information from the edge types in order to determine the communities. Fundamental to our approach is the *relation*, which we define as all edges between two vertices. We build upon recent advancements in edge-based community detection [4, 1, 2] by directly incorporating edge type information into our definition of community quality.

Because information conveyed by the edge types will vary from network to network, we introduce a new function σ over the types that defines their similarity. For two edge types, t_i, t_j, let $\sigma(t_i, t_j) \rightarrow [-1, 1]$ be the value indicating the types' similarities. While σ may be defined by the investigator based on *a priori* knowledge of the network, σ

Cross and Parker (CP) [3] Network Summary
Cross and Parker interviewed employees at a European manufacturing research firm, located in four cities. Interviews were done for individuals in different organizational positions and at different levels of tenure at the company. Individuals were asked to rate other individuals on a Likert scale from 1–6 according to two types of social knowledge: (1) the degree to which they use information from other person to accomplish their work, and (2) their awareness of the other person's skills and knowledge. Responses of 1 indicate that the individual does not know the other person.
Network Construction Each individual is represented as a node. For each question, a directed edge is added from individual i to j if i answered above 1 on the scale, indicating some level of familiarity with j. Furthermore, each answer in the scale was treated as a separate edge type, e.g., an awareness level of 2 is a different edge type from an awareness level of 5. The resulting network has 10 edge types, with 77 individuals and 3901 edges.

Table 1: A description of the multiplex social network gathered by Cross and Parker.

could also be defined using external ontologies to compare relation labels, or through comparing labels with data-driven approaches such as distributional semantics. For example, in a social network modeling with edge types represent communication topics, the type similarity could be defined according to the terminology overlap between the topics, thereby avoiding manual specification.

Relation Comparison

The core of our method is based on the definition of edge similarity, which is subsequently used to identify edge partitions that give rise to communities. As a starting definition for edge similarity, we build upon the definition of Ahn et al. [1] which was defined for undirected, untyped networks. For two edges, $e_{i,k},\ e_{j,k}$, their similarity is defined using the Jaccard Index of the set of inclusive neighbors:

$$sim(e_{i,k},\ e_{j,k}) = \frac{N_i^+ \cap N_j^+}{N_i^+ \cup N_j^+} \qquad (1)$$

where N_i^+ is the set containing vertex i and its neighbors. This definition successfully captures the intuition that two edges are similar if they share the majority of their neighbors. However, a multiplex network allows for a richer comparison of edge similarity that takes into account both the number of connections to a node, turning the set of neighbors into a multiset, as well at the types of those connections.

To extend Eq. 1, we note the expectation that two vertices are more likely to be in the same community if the neighbors they have in common are connected by similar relations. Therefore, we redefine Eq. 1 to capture both the similarity of the edge types as well as the notion of connectivity in a multiplex network. Let $\mathbb{N}_v$ be the multiset of vertices connected to vertex v and $R_{i,j}$ be the set of edges connecting vertices i and j. The similarity of two relations is defined as:

$$sim(R_{i,k}, R_{j,k}) = \frac{f(\mathbb{N}_i, \mathbb{N}_j)}{|\mathbb{N}_i|^2 + |\mathbb{N}_j|^2 - f(\mathbb{N}_i, \mathbb{N}_j)} \qquad (2)$$

where $f(\mathbb{N}_i, \mathbb{N}_j)$ measures the degree of similarity between the edges sets:

$$f(\mathbb{N}_i, \mathbb{N}_j) = \sum_{z \in \mathbb{N}_i \cap \mathbb{N}_j} \sum_{e_{i,z} \in R_{i,z}} \sum_{e_{j,z} \in R_{j,z}} \sigma(e_{i,z}, e_{j,z})$$

Note that if σ=1 for identical edge types and 0 for all others, Eq. 2 simplifies to the Tanimoto Similarity, which is a generalization of the Jaccard Index used in Eq. 1 for calculating the similarity of multisets. The relation similarity defined in Eq. 2 is proportional to the percentage of neighbors that are shared in common with respect to the similarity of the edge sets connecting the neighbors. Eq. 2 is maximized when i and j have identical neighbors and the edges to those neighbors have maximal similarity; and conversely, it is minimized when when i and j have identical neighbors but those neighbors have negative similarity, which indicates that i and j have opposing relations to all their neighbors.

Clustering Relations and Identifying Communities

The relation similarity function in Eq. 2 allows us to rate relationship according to how likely their corresponding vertices should be in a community together. We adopt a two-phase approach for community detection: (1) multiple clustering solutions are generated at different community granularities, and (2) solutions are evaluated using a criterion function for intra-community connectivity and similarity, ultimately selecting the solution with the highest value. We first discuss how community solutions are evaluated and then describe how they are generated.

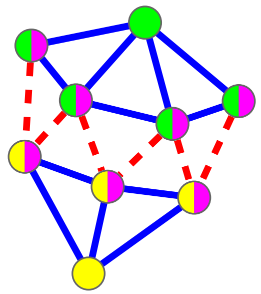

Figure 2: The community membership of the friend-enemy network from Fig. 1 using our agglomerative clustering method. The solution identifies three communities: two consisting of only friends, and a third community made of individuals who are enemies of each other. Community membership is shown by colors in the vertices' symbols. In contrast, the purely edge-based solution of [1] merges the network into a single community.

An ideal individual multiplex community maximizes the number of edges between its vertices, while simultaneously maximizing the average similarity of the edge types in the community. We build upon the community density function of Ahn et al. [1] to define a new criterion that incorporates both density and semantic similarity. Let n be the number of vertices in a community and m be the number of edges between the members of that community. For a multiplex network with t types, the maximum number of edges in a community is $\frac{t \times n(n+1)}{2}$, while the minimum is $n-1$. Therefore, the density of a community in terms of its edges is computed as

$$d(C) = \frac{m-(n-1)}{\frac{t \times n(n-1)}{2} - (n-1)} \tag{3}$$

where d=0 if n=2.

We define the similarity of edges in community C as:

$$s(C) = \sum_{e_i, e_j \in C | i \neq j} \sigma(e_i, e_j) \tag{4}$$

Therefore, we measure the total goodness of the community solution with k communities as the sum of each community's density relative to the total number of edges in the network and weighted by the global similarity of relations within the communities:

$$\frac{2}{kM \sum_{i=1}^{k} \frac{|C_i|(|C_i|-1)}{2}} \sum_{i=1}^{k} d(C_i)s(C_i) \tag{5}$$

By weighting the density according the clustering similarity, this equation penalizes structurally consistent communities that have highly dissimilar edge types connecting their vertices. We normalize by the global intra-community similarity rather than the average community similarity in order to take into account the difference in community sizes so that a large self-similar community has more weight than a small dissimilar community.

The final community solution is computed by creating communities from each of the relation clusters, where a community contains all of the vertices connected by relations in that cluster. Relations are clustered using hierarchical agglomerative clustering, which continuously merges the most similar clusters according to a criteria function in order to build a dendrogram of how all relations are connected. We adopt the computationally-efficient strategy of Ahn et al. [1] and use the single-link criterion, which merges the two clusters containing the most similar relations according to Eq. 2 that are not currently in a cluster together. The agglomerative clustering produces a dendrogram over all relations, which is cut at each level to produce a community solution that is evaluated according to Eq. 5. The solution that maximizes Eq. 5 is selected as the results. For a network with R relations, our method requires only $O(R^2)$ time, which is sufficient to easily scale to networks with hundreds of thousands of edges.

Experiments

As an initial study, we consider both the synthetic friend-enemy network from Fig.1 and the real-world social network of researcher gathered by Cross and Parker (CP) [3], described in Tab. 1.

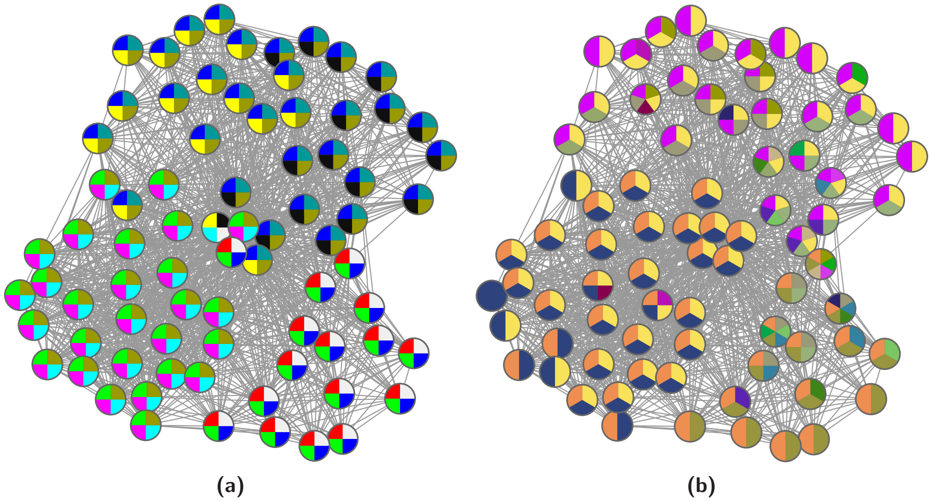

Figure 3: Figure 3a highlights the 10 communities found in the CP network using our method. Figure 3b shows the 20 communities found using the method of Ahn et al. [1]. Each vertex color corresponds to a specific community membership. For simplicity, parallel edges are visualized as a single edge.

Central to our method is the selection of the σ function. Therefore, in the friend-enemy network, we considered a range of definitions for σ

$$\sigma(t_1, t_2) = \begin{cases} t_1 = t_2 & 1 \\ t_1 \neq t_2 & x \end{cases}$$

where x defines the similarity of non-equal types. Figure 2 shows the resulting communities when x=0, which match the expectation that friends and enemies be segregated into separate communities. Furthermore, our analysis showed that the method is not sensitive to the value of x, with all solutions having $-1 \leq x < 0.8$ generating the same community divisions.

The CP network contains edges types in two sets T_1 and T_2, reflecting the two different social relations. We define σ by noting that each type t_i can be associated with the weight of the relation $w(t_i)$ as defined in the original Cross and Parker survey. Comparisons between types in the same set may be made according to the relative differences in their weights. Therefore we define σ as:

$$\sigma(t_1, t_2) = \begin{cases} t_1 \in T_i \wedge t_2 \in T_i & 1 - \frac{|w(t_1) - w(t_1)|}{5} \\ t_1 \in T_i \wedge t_2 \in T_{j \neq i} & 0 \end{cases}$$

Fig. 3 shows the resulting communities, in comparison with those found by the method of Ahn et al. [1].

The communities found by our method reveal two notable patterns. First, communities strongly correspond to the geographic locations of individuals in the CP network, detailed in Table 2. Six of the communities correspond to geographically-homogeneous individuals. An analysis of the remaining four geographically-heterogeneous communities revealed each was connected by a single individual who acted as a bridge between the two or more geographic areas, with multiple individuals from each location being familiar with the person. These bridge individuals appear in the center of network shown in Fig. 3a. In contrast, the purely link based approach of Ahn et al. [1] finds communities that roughly correspond to two of the locations and a large community that blends the remaining location. Our analysis did not reveal in further connection from individuals' tenure or organization level to the communities in their solution, despite having identified twice as many communities. Overall, we view our method as providing an insightful exploratory tool in identifying salient structure of the network from its multiplex relations alone.

Com.	Pa.	Fr.	Wa.	Ge.
1	-	-	-	1.00
2	-	0.63	-	0.37
3	0.35	-	0.33	0.33
4	-	1.00	-	-
5	-	-	0.94	0.06
6	-	0.96	-	0.04
7	-	-	-	1.00
8	0.94	-	-	0.06
9	0.28	0.45	0.27	-
10	0.52	-	0.48	-

Table 2: The percentage of individuals in each community from Fig. 3a located in cites Paris, Frankfurt, Warsaw, or Geneva.

Conclusion and Future Work

We have raised the issue of how to discover communities in multiplex networks where the edge type convey important information on the different types of relationships. Accordingly, we have proposed a new approach to community detection based on an edge type similarity function σ. Our initial results suggest that this method can identify communities that meet expectations of which entities should be grouped.

Furthermore, our results raise several significant questions for future work. First, community detection is frequently evaluated in terms of network modularity [9], which tests the community partitioning relative to a null model. However, many modularity definitions are defined in terms of binary, weighted, or directed connections, and therefore do not take into account the parallel edges present in multiplex networks. Future work is needed to define modularity in multiplex networks to quantify the goodness of a particular solution. Second, the present work has only considered edge types. However, many networks contain distinct vertex types, or additional vertex properties. Future work is also needed to assess how these properties can be incorporated into our definitions of community quality. Last, future work is needed to establish standardized mulitplex network datasets and corresponding community soluations to enable better comparisons between approaches.

References

[1] Y. Ahn, J. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

[2] B. Ball, B. Karrer, and M. E. J. Newman. Efficient and principled method for detecting communities in networks. *Physical Review E*, 84:036103, Sep 2011.

[3] R. Cross and A. Parker. *The Hidden Power of Social Networks: Understanding how work really gets done in organizations*. Harvard Business School Press, 2004.

[4] T. Evans and R. Lambiotte. Line graphs, link partitions, and overlapping communities. *Physical Review E*, 80(1):016105, 2009.

[5] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[6] D. Greene and P. Cunningham. Multi-view clustering for mining heterogeneous social network data. In *Workshop on Information Retrieval over Social Networks, Proc. of ECIR*, 2009.

[7] A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. *Phys. Rev. E*, 80:056117, Nov 2009.

[8] P. Mucha, T. Richardson, K. Macon, M. Porter, and J. Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876, 2010.

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

[10] M. Rodriguez and J. Shinavier. Exposing multi-relational networks to single-relational network analysis algorithms. *Journal of Informetrics*, 4(1):29–41, 2010.

[11] M. Szell, R. Lambiotte, and S. Thurner. Multirelational organization of large-scale social networks in an online world. *PNAS*, 107(31):13636, 2010.