Detecting Communities of Interests in Social Media Platforms using Genetic Algorithms

Full Paper

Mahyar Sharif Vaghefi
University of Wisconsin-Milwaukee
mahyar@uwm.edu

Derek L. Nazareth
University of Wisconsin-Milwaukee
derek@uwm.edu

Abstract

Detecting communities of interest in social media platforms provides insight into the platforms and the individuals that use them. The bulk of research in community detection is directed at network analysis of individuals and their interaction with other members within the network. However, connections outside the network can also be useful for community detection, as in the following of elite Twitter users by regular users. This research develops a mechanism for clustering elite Twitter users on the basis of connections and interactions within their followers. Since clustering is sensitive to initial configurations, the approach is modified using genetic algorithms to traverse multiple regions of the solution space. Application of this approach to a set of 25,000 Twitter users demonstrates that it forms coherent communities within a few iterations, outperforming other clustering approaches for community detection.

Keywords

Community Detection, Social Media, Genetic Algorithm, Twitterati, Communities of Interest

Introduction

There has been a tremendous growth in online social network sites recently. According to the Pew Research Center, almost two-third of American adults use social networking sites in their daily activities. This has revolutionized information sharing and communication patterns, and affected areas as work and politics (Perrin 2015), spawning new manners of community formation. Ren et al. (2007) defined online community as "an Internet-connected collective of people who interact over time around a shared purpose, interest, or need". In their study of online communities, Papadopoulos et al. (2012) classified these as explicit and implicit communities. Explicit communities involve a conscious decision to be part of a group, while implicit communities are formed by day to day interactions of individuals and are not always visible to all. Identifying communities in social networks permits monitoring and providing services to them. Outside interested parties could include marketers seeking to provide targeted ads, law enforcement seeking to monitor terrorists, and social network administrators looking to protect members.

Several approaches have been adopted to identify communities in social networks. Some of these use the structure of user connections in the network to find groups that are more densely connected to each other than with the rest of the network. Papadopoulos et al. (2012) provide a good overview of different structural based algorithms and their application in social media. Another approach focuses on the users' interactions and their pattern of communication within the network. This goes beyond the connections and examines patterns of interactions, e.g. comments and tweets (Palsetia et al. 2012; Deitrick and Hu 2013), expanding network analysis algorithms to incorporate the dynamic interaction to form and detect communities. A third approach goes even deeper, looking to extract meaning through the use of topic selection techniques and semantic analysis (Zhao et al. 2012; Xia and Bu 2012).

These approaches work well for forming communities based on direct information about users, e.g. connections, interactions, and content dissemination. Thus for example, it is easy to form communities of Twitter users based on who they follow, message, their retweets, or hashtag usage. However, Twitter represents a platform where elite users can influence several other followers. Clustering elite Twitter users (or Twitterati) based on their followers and their interactions with other Twitter users represents a

new opportunity for research, wherein implicit commonalities can be detected through secondary interactions, based on the concept of homophily (McPherson et al. 2001). This research looks to form Twitterati clusters, not by examining their direct connections, but through the interactions of their followers. This constitutes the clustering of a bipartite graph. Using a data set collected from Twitter, it builds clusters of influential Twitter users in an effort to better understand their implicit relationships based on their followers.

The rest of the paper is organized as follows. The next section provides a brief overview of the graph clustering literature and discusses current clustering approaches in social media networks. It highlights the limitations of current approaches for clustering a bipartite graph, and creates the case for the use of an evolutionary algorithm to facilitate community detection in this context. Details of the proposed algorithm are presented in the following section. The approach is applied to a large Twitter data set, and compared with other clustering approaches. Implications for this approach round out the paper.

Graph Clustering

Literature Review

The primary objective in graph clustering is to partition a graph into meaningful parts in a manner that maximizes the proportion of inter-cluster edges to intra-cluster edges (Schaeffer 2007). A graph cluster is often considered to be a community (Girvan and Newman 2002). Well-formed communities are "cohesive, compact and internally well connected while being also well separated from the rest of the network" (Yang and Leskovec 2015).

Several measures have been proposed to measure the goodness of a community. These are classified into four broad types of measures based on their focus – internal connectivity, external connectivity, internal and external connectivity, and network level (Yang and Leskovek 2015). As with many measures that seek to distill the essence of a collection, there is substantial difference and variance observed in the measures. Table 1 provides an overview of the prominent measures used in this community assessment literature.

Type	Measure	Researcher	Description
Internal	Density	Radicchi et al. (2004)	Measures how well members of a community are connected to each other, by dividing the number of existing edges among the members of a community by the number of all possible connections among them.
	Average Degree	Radicchi et al. (2004)	Represents the average number of connection for each node inside a community.
	Edge Count	Radicchi et al. (2004)	Counts the number of edges in a community.
External	Cut Ratio	Fortunato (2010)	Proportion of outside community connections to the all possible number of connections to the outside community.
	Expansion	Radicchi et al. (2004)	Average number of outside community connection for each node inside the community.
Internal & External	Conductance	Leskovec et al. (2009)	The fraction of outside community edges to the sum of the degree of inside community nodes.
	Normalized Cut	Shi and Malik (2000)	Fraction of outside community edges to the degree of inside community nodes plus fraction of outside edges to the degree of outside community nodes.
Network	Modularity	Newman and Girvan (2004)	Number of connections within the community relative to expected value in the random graph.

Table 1: Prominent Measures for Assessing Goodness of a Community

Though the measures have been created for post-hoc assessment of existing communities, recently they have been applied to community detection (Fortunato 2010). Algorithms for detecting communities have been characterized as overlapping or non-overlapping, based on the nature of communities detected. The latter algorithms refer to communities that are distinct wherein each node is assigned to one community only, as described in (Blondel et al. 2008), (Clauset et al. 2004). In reality, nodes can be shared among different communities leading to overlapping areas. Several algorithms have been developed to address the overlapping community detection (Yang and Leskovec 2013), (Yang et al. 2013).

Graph clustering algorithms for community detection have been used occasionally in the study of social media populations. Feng et al. (2015) developed an overlapping community detection algorithm to find different interest based communities in online social networks sites, using datasets from MovieLens and Netflix. In another study, Gonzales-Bailon and Wang (2016) applied community detection algorithms to study protest campaigns in social media and showed that global brokerage positions in online social networks can play more important roles than local brokerage positions. Papadopoulos et al. (2012) catalogued community detection algorithms for social media into five categories based on community definition and underlying methodology. Using a set of six community formation primitives, they examined the Lycos iQ question forum, and used seed tags to detect communities. Yang and Leskovec (2015) stressed the need for functional properties in implicit communities whereby community members would share common properties including affiliations, roles or attributes. They argued that implicit community detection becomes a two-phase process – community detection based on network structure, and confirmation through common attribute identification. The process of discovering the functional property of a network may require considerable effort. Therefore, in this study we offer a new method for clustering users' interests into distinct groups that can later be used as an appropriate functional property.

Motivation for bipartite graph clustering

In most analyses of communities in social networks, researchers limit the network to a set of users and examine the network structure among users, ignoring connections they may have to other nodes outside the set selected. Consequently, useful semantic information is ignored in this approach to community detection. Incorporating this information would result in the formation of an extended bipartite graph. In the Twitter context, these would constitute Twitter users and Twitter elite, and is illustrated in Figure 1.

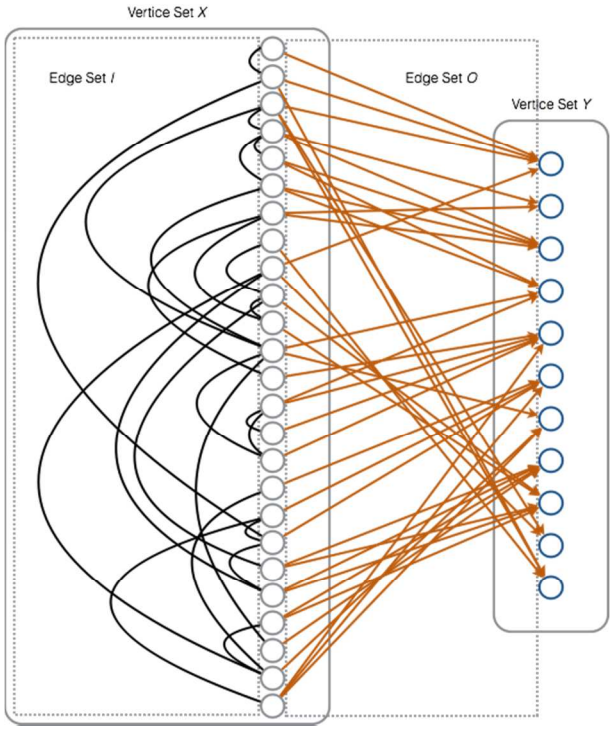

Figure 1. Bipartite Graph of Users

In a bipartite graph, vertices belong to one of two classes X and Y, where X∪Y=V and X∩Y= ∅. Likewise, there are two edge sets, I and O, where each edge in I has two endpoints in X, while each edge in O has one endpoint in X and one endpoint in Y, such that I∪O=E and I∩O= ∅.

In order to cluster Twitterati based on their followers, we need to devise an algorithm for clustering vertices in Y, based on the vertices in X and their connections in I. This would essentially generate topics communities of interest for users in X. Given that both X and Y can be large sets in the Twitter context, specifying the number of clusters becomes critical. A very small number leads to diffuse clusters, while a large number generates small and less meaningful clusters. The judicious selection of the number of clusters allows for Twitter users to coalesce around meaningful topics and themes, and facilitates the functioning of the algorithm since it now uses meaningful in-degree and out-degree values.

Similarity-Based Clustering Algorithm

We now proceed to a discussion of our proposed algorithm. Let $X = \{x_i \mid 1 \le i \le n_x\}$ be the set of labeled nodes that represents the population of interest in online social network and $Y = \{y_i \mid 1 \le i \le n_y\}$ be the set of labeled nodes outside the network that are followed by nodes in X, where n_x and n_y be the number of users in the population of interest and number of users outside the user network respectively. We can model this network using two adjacency matrices $A=\{(V_i,V_j) | i \neq j \text{ and } V_i,V_j \in X\}$ and $B=\{(V_i,V_j) | V_i \in X, V_j \in Y\}$. As a result, each node in Y can be represented by a labeled graph of nodes in X, as illustrated in Figure 2.

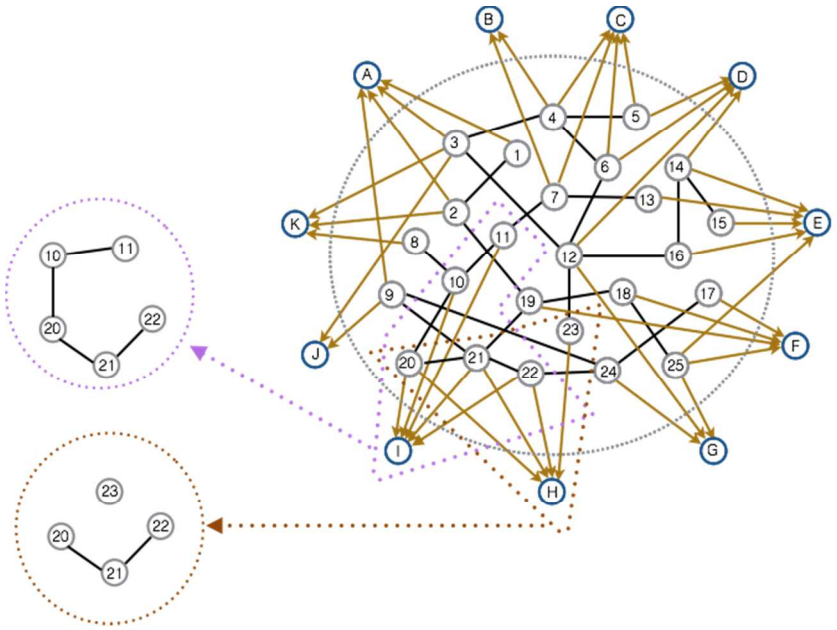

Figure 2. Mapped Graphs

Similarities in a graph can serve as a basis for clustering a mapped graph, i.e. two users in Y with similar set of followers in X and similar pattern of connections among them are more similar other than two users with diverse set of followers. This is consistent with the concept of homophily, which posits that connections between similar people occur more frequently than dissimilar people (McPherson et al. 2001). Similarity of mapped graphs is computed using the following:

$$Sim(G_1, G_2) = \frac{(|V(G_1, G_2)| + |E(G_1, G_2)|)^2}{(|V(G_1)| + |E(G_1)|) \,.\, (|V(G_2)| + |E(G_2)|)} \quad (1)$$

where $|V(G)|$, $|E(G)|$ are the number of vertices and number of edges in a graph, $|V(G11,G2)|$ is the number of common vertices between graph $G1$ and $G2$, and $|E(G1,G2)|$ is the number of common edges between graph $G1$ and $G2$. The similarity measure returns a value between 0 and 1, where 0 indicates no similarity and 1 indicates complete similarity. For instance, the similarity value of two graphs in Figure 2 is 25/54, or 0.46. The properties of this measure are documented in (Johnson 1985). Our algorithm uses

mapped graph concepts and the similarity measure to cluster the vertices, and is outlined in Table 2. It requires the identification of cluster cores, which is illustrated in Figure 3.

1. Determine the desired number of clusters K, and randomly assign each of the vertices in Y to one of the clusters. 2. Find the core of the clusters $Z=\{z_i \mid 1 \le i \le k\}$. The core z_i is defined as the segment that is common to a number mapped graphs specified by a threshold t. 3. Find the similarity of all mapped graphs and the core of the clusters and reassign the mapped graphs to the clusters based on the measure of similarity. 4. If the assignment of mapped graphs to clusters in step 3 leads to a new composition of clusters, repeat set 2, else terminate.

Table 2. Proposed Clustering Algorithm

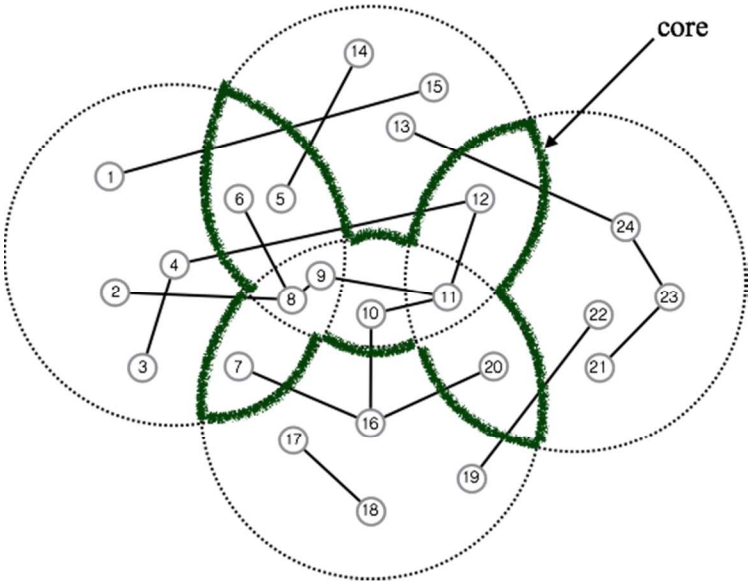

Figure 3. Cluster Core Identification with Threshold of 2

The clustering approach outlined in Table 2 is rather sensitive to the initial random assignment of nodes to clusters. As a result, an unfavorable initial assignment will yield poor results. In order to improve the final clustering we employ a genetic algorithm. This is an evolutionary heuristic approach that employs ideas from population genetics and works effectively with large problem spaces (Holland 1975).

Clustering Using Genetic Algorithms

Genetic algorithms (GA) employ a set of population-based operators to purposefully navigate through a problem space to find the best solutions. They have been widely employed, ranging from applications in biology, computer science, information technology, logistics, economics, power systems, and production and operations management, among others (Beasley et al. 1993). They have been adapted to address the problem of K-means clustering (Maulik and Bandyopadhyay 2000). This adaptation involves a restructuring of the GA approach to some extent. We employ these notions to permit the clustering algorithm to search different regions in the solution space and avoid local optima regions. In the adapted GA, we seek to find the set of cores z_i that maximizes the similarity across all clusters.

$$\sum_{i=1}^{k} \sum_{G_j \epsilon C_i} Sim(G_j, z_i) \tag{2}$$

Genetic algorithms employ a population of individuals each representing a solution that is encoded as a binary chromosome. A fitness function is employed to select the best individuals, and a set of population generation transformations, like crossover and mutation, are applied to generate new solutions. A

stopping rule is employed to terminate the process. Unlike a standard GA, where a chromosome represents a complete solution, in this case each chromosome comprises k alleles that represent the vertices of a clusters' core. This is illustrated in Figure 4 assuming a threshold t of 2. In this case, the binary vector in each allele is used in conjunction with the overall network adjacency matrix to find the core graph of each cluster. The initial population consists of p individuals. Each individual is represented by a chromosome comprising k alleles. The chromosome is generated by random assignment of vertices in Y to clusters, in a manner such that they result in a total of k cluster cores, and are given by $Z=\{z_i | 1 \leq i \leq k\}$.

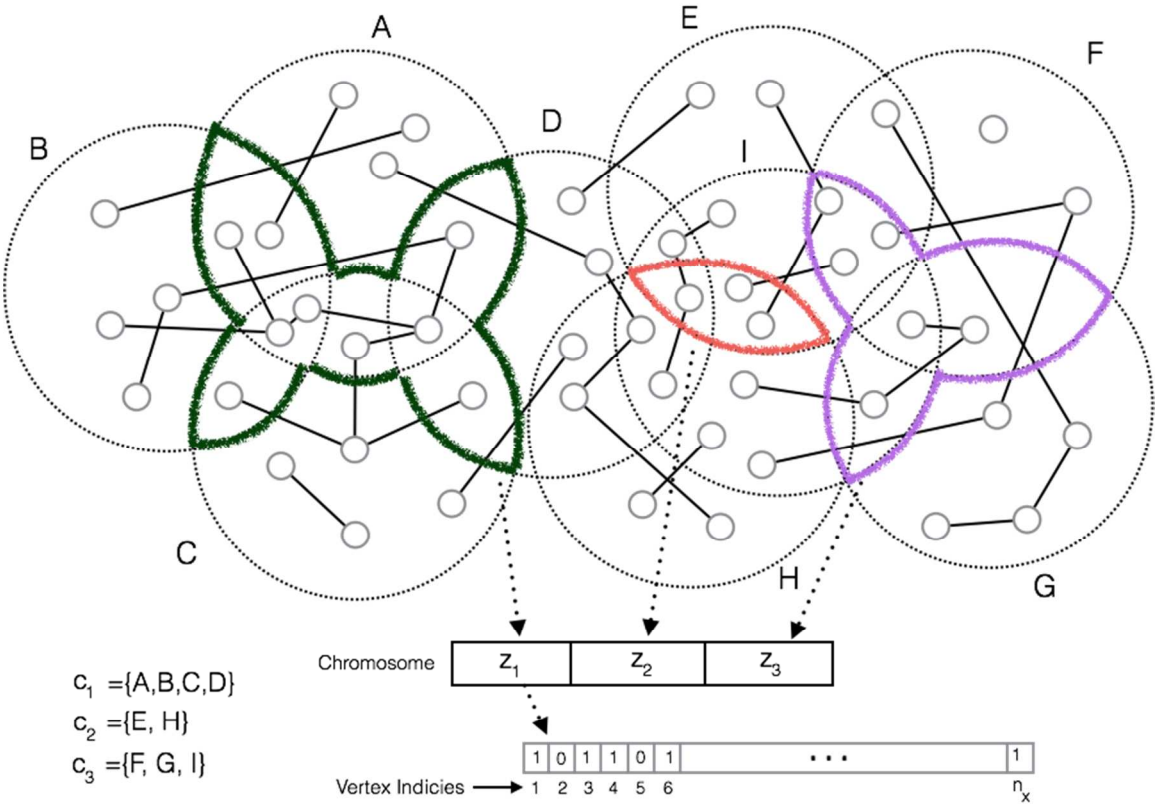

Figure 4: Chromosome Representation

The fitness computation employed in this approach comprises four steps. First, the similarity of the mapped graph of each vertex in Y and the core of each cluster in chromosome p_i is computed using the Equation 1. Second, vertices in Y are reassigned to the clusters based on their similarity values. Third, the core of each cluster is updated based on the new configuration. Finally, the fitness value of each chromosome is computed using the Equation 2. An elite selection method is used in our GA formulation. We use the top 50% of the population to generate new offspring for the next generation of the GA. The offspring are generated through a combination of crossover and mutation functions. Crossover represents a transformation that combines the front and rear portions of two parent chromosomes to generate an offspring chromosome. Our algorithm uses a single point crossover with a random allele crossover point. In addition, to avoid having the GA trapped in a local optima region, we also apply a mutation operation with a fixed probability p_m. If the probability exceeds the mutation threshold, a random bit is flipped in the chromosome, generating a slightly different solution. Termination criteria in GAs include attainment of a specific value, lack of improvement in successive iterations, or a fixed number of iterations. In this case, we used a fixed number of iterations, restricting it to 30, since there was little improvement noted beyond this.

Data Collection

We used the Twitter API to collect the data used to evaluate our approach. The initial population of

interest was identified from tweets that mentioned a Foursquare check-in that were issued between July 8, 2015 and July 28, 2015. Our aim was to identify users engaged in real activities. This provided us with online profiles for 27,167 users. The users' profiles contain basic information about individuals including such as the number of tweets, number of friends and followers, and a brief description of the individual.

We used the Twitter API once again to collect information about the social network connections of these individuals. This allowed us to establish two sets of connections – connections among individuals in the population, and connections to individuals outside the population, as indicated in the bipartite graph. The former set consists of several unconnected networks, the largest portion of which comprises more than 85% of all individuals. We focus on this set, and consider only cases of two-way connections between individuals, since these are considered more stable than one-way connections (Kwak et al. 2011). Dropping the one-way connections reduces the number of edges by 8%, leaving us with a strongly connected network of 25,309 nodes and 93,682 edges. For the second set, we focused on the top 100 to 200 (N) Twitter accounts that are followed by the 25,309 users, thereby setting up the bipartite graph for community detection through Twitter elite clustering.

Clustering Algorithm Evaluation

We evaluated the clustering algorithm on the selected data set using several clustering parameters. These included the number of different clusters (K=3,4,5), the threshold value for determining the core of each cluster ((T=2,3,4), and the number of Twitter elite users for clustering (N=100, 200). Since the eventual clustering is dependent on the initial starting point, we ran each combination five times. We also used a stopping rule of 30 iterations. In all cases, the fitness value stabilized long before the end point was reached. The results indicated that increasing the number of clusters (K), and the threshold value (T), led to higher values of the clustering objective value.

We then added the GA component to examine if the clustering results could be improved through a more widespread search. We used the same value of parameters (K, N, T) that we used for clustering algorithm. For the GA, we set the number of chromosomes to 10, and employed the same stopping rule of 30 iterations. The fitness value stabilized before the end point, and the GA-based approach continually outperformed the clustering approach, with a final improvement of about 7%. This indicates that the clustering approach is dependent on the initial assignment of individuals to clusters. While the results are very encouraging from a measure of clustering goodness perspective, we wanted to examine the results from a semantic perspective. We used the ForceAtlas 2 layout (Jacomy et al. 2014) in Gephi (Bastian et al. 2009) to help visualize the clustered network, as illustrated in Figure 5. The bright nodes demonstrate the top 100 Twitter accounts that are followed by 25,309 users in our population of interest.

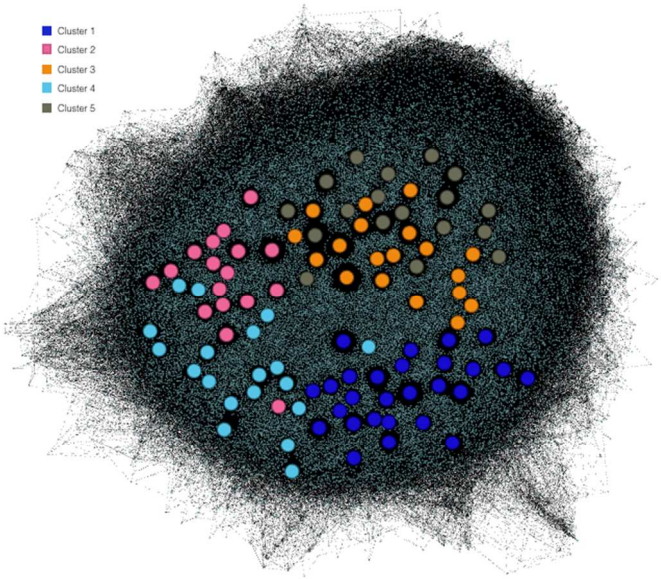

Figure 5. Clustered Network of 100 Twitter Elite

Figure 5 clearly demonstrates that individuals in online social networks belong to different communities of interest. We then examined the assigned Twitter accounts in each cluster. These results are depicted in Table 3, with one account deactivated by its owner. As with any clustering technique, some outliers are to be expected. Nonetheless, the communities of interest are quite evident in the resulting clusters. We termed these communities Comedy and Satire, Music and Television, News and Politics, Sports and Entertainment, and Commercial and Technology. The effectiveness of clustering is also dependent on the number of clusters – too few results in clusters that are not cohesive, while too many leads to fragmented and meaningless clusters.

Cluster 1	Comedy and Satire	Wil Wheaton, Aziz Ansari, The Onion, Anthony Bourdain, Joel McHale, Steve Martin, Jimmy Fallon, Funny Or Die, Stephen Colbert, Seth MacFarlane, RainnWilson, Bill Maher, Neil deGrasse Tyson, {Deactivated account}, Mindy Kaling, Joseph Gordon-Levitt, Daniel Tosh, Jimmy Kimmel, Sarah Silverman, Zach Galifianakis, Seth Meyers, Tom Hanks, Justin Halpern, Zooey Deschanel, Neil Patrick Harris, Conan O'Brien, The Daily Show
Cluster 2	Music and Television	Lady Gaga, Ellen DeGeneres, Ryan Seacrest, Britney Spears, Taylor Swift, Ashton Kutcher, Perez Hilton, Oprah Winfrey, Chelsea Handler, Katy Perry, Kim Kardashian West, Justin Timberlake, P!nk, Jim Carrey, Adam Levine, Rihanna, Adele
Cluster 3	News and Politics	CNN Breaking News, CNN, The New York Times, Barack Obama, Wall Street Journal, NPR, BuzzFeed, Breaking News, NASA, TIME.com, Huffington Post, Anderson Cooper, Rachel Maddow , Dalai Lama, The White House, The Associated Press, Michelle Obama, Bill Clinton, Hillary Clinton
Cluster 4	Sports and Entertainment	ESPN, Mark Cuban, TMZ, SHAQ, NFL, TextsFromLastNight, Marshall Mathers, LeBron James, Kevin Hart, Rev Run, SportsCenter, Drizzy, Adam Schefter, OMG Facts, UberFacts, Kanye West, Dwayne Johnson, Jonah Hill, Charlie Sheen
Cluster 5	Commercial and Technology	Starbucks Coffee, Twitter, TechCrunch, Mashable, WIRED, Southwest Airlines, YouTube, TwitPic, Foursquare, Dropbox, Whole Foods Market, Klout, Hootsuite, Google, Bill Gates, Pinterest, Instagram, Vine

Table 3: Cluster Members

Inspection of the clusters suggests that the results are fairly intuitive. However, none of the semantic information about the domains that these Twitterati are affiliated with was used to generate these clusters. Rather the commonality of their followers was exploited to cluster them. This indicates the power of this approach, wherein community detection in the bipartite network yield hidden meaning not ordinarily present in the network.

We also examined the composition of the clusters generated in each iteration. Variation was observed in the cluster cores, indicating that the GA was making a difference in formation of clusters, indicating that areas of the solution space were being explored. The results indicate that clustering of Twitter elite using interconnectedness of their followers can yield clear and meaningful communities of interest.

Implications

Prior research in community detection in social networks has focused on the use of network properties. By incorporating information outside the network (elite Twitter users in this case), we are able to discern communities that incorporate additional semantics. To further improve the clustering approach, this study adopts a genetic algorithm scaffold for the clustering, so as to reduce dependence on initial assignment.

This paper makes several contributions to the clustering of nodes in a large and densely connected network. First, it uses external information to develop clusters, which represents a novel contribution.

The use of external information provides additional semantics that permit the formation of meaningful clusters. Second, by using the overlap between clusters, we are able to identify cluster cores, which are instrumental in forming cohesive clusters. Incorporation of genetic algorithm principles permits us to search larger portions of the space, thereby reducing the probability of being trapped in a local optima region. The resulting communities of interest are shaped by external information in a manner that yields cohesive clusters. Comparison with other network clustering techniques indicates that our approach performs better on standard similarity measures.

The technique was applied to elite Twitter users to detect communities of interest. Classifying the elite users in social networks based on interactions of their followers unearths additional information about the communities of interest that is not ordinarily available in the standard network measures. The fact that the algorithm resulted in clusters that are considered intuitive, indicates the power of bipartite network clustering. Analysis of this information can help professionals who study social networks. Developers of recommendation systems can capture these communities of interest to enhance the effectiveness of their recommendation systems. Marketing professionals can use these findings to better target customers for their promotional offers. Law enforcement can use these techniques to identify individuals and groups that warrant additional attention, by studying their followers. Cultural studies can also benefit by helping unearth communities of individuals that share or have interest in a particular culture, even though explicit articulation of shared interests is not available. Any analysis of a social network that seeks to identify communities, particular in the presence of external influences can benefit from the application of these techniques.

Conclusions

In this study, we used a novel approach to introduce a graph clustering algorithm that clusters nodes outside a social network on the basis of properties within the network. Our algorithm is based on the similarity of graph structure of the followers inside the network. The application of this method is not limited to social media networks and can be used for clustering of all extended type of bipartite graphs that contain additional information about the structure of connections among the nodes. For instance, this algorithm can be used to classify the different social events by having information about the social network connection of the participants in these events.

The results indicate that the clustering algorithm is effective, both in terms of the goodness of the resulting clusters as measured by standard graph clustering measures, as well as soundness of the clusters as determined through an examination of the cluster semantics. Incorporation of a genetic algorithm structure further enhances the effectiveness of this approach, by permitting search of a larger portion of the space, and thereby reducing the dependence on the initial assignment of clustering, while also resulting in more meaningful clusters.

Acknowledgments

The first author was supported by a fellowship from the UWM Graduate School.

References

Bastian, M., Heymann, S., and Jacomy, M., 2009. “Gephi: An Open Source Software for Exploring and Manipulating Networks,” in *International AAAI Conference on Web and Social Media* (*ICWSM)*, pp. 361-362.

Beasley, D. Bully, D.R., and Martinz, R.R., 1993. "An Overview of Genetic Algorithms: Part 1, Fundamentals, *University Computing*, (15:2), pp 58-69.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. 2008. "Fast Unfolding of Communities in Large Networks", *Journal of Statistical Mechanics: Theory and Experiment*, (2008:10), p. P10008.

Clauset, A. Newman, M.E.J. and Moore, C. 2004. "Finding Community Structure in Very Large Networks." *Physical Review,* (70:6), p 066111.

Deitrick, W. and Hu, W. 2013. “Mutually Enhancing Community Detection and Sentiment Analysis on Twitter Networks”, *Journal of Data Analysis and Information Processing*, (1:3), pp.19-29.

Feng, H., Tian, J., Wang, H.J. and Li, M. 2015. “Personalized Recommendations Based on Time-Weighted Overlapping Community Detection”, *Information & Management*, (52:7), pp.789-800.

Fortunato, S., 2010. "Community Detection in Graphs," *Physics Reports*, (486:3), pp. 75-174.

Girvan, M. and Newman, M.E., 2002. "Community Structure in Social and Biological Networks," in *Proceedings of the national academy of sciences*, (99:12), pp.7821-7826.

González-Bailón, S. and Wang, N., 2016. "Networked Discontent: The Anatomy of Protest Campaigns in Social Media," *Social Networks*, (44), pp. 95-104.

Holland, J.H. 1975. *Adaptation in Natural and Artificial Systems*, Cambridge, MA: MIT Press.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M., 2014. "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software," *PloS one*, (9:6), p.e98679.

Johnson, M., 1985. "Relating Metrics, Lines and Variables Defined on Graphs to Problems in Medicinal Chemistry," in *Graph Theory with Applications to Algorithms and Computer Science,* Y. Alavi, G. Chartrand, L. Lesniak-Foster, D. R. Lick, and C. E. Wall (eds.), New York: Wiley Interscience, pp. 457-470.

Kwak, H., Lee, C., Park, H. and Moon, S., 2010. "What is Twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web*, pp. 591-600.

Leskovec, J., Lang, K.J., Dasgupta, A. and Mahoney, M.W., 2009. "Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters," *Internet Mathematics*, (6:1), pp. 29-123.

Maulik, U. and Bandyopadhyay, S., 2000. "Genetic Algorithm-Based Clustering Technique," *Pattern Recognition*, (33:9), pp. 1455-1465.

McPherson, M., Smith-Lovin, L. and Cook, J.M., 2001. "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, (27), pp. 415-444.

Newman, M.E. and Girvan, M., 2004. "Finding and Evaluating Community Structure in Networks," *Physical Review E*, (69:2), p.026113.

Palsetia, D., Patwary, M.M.A., Zhang, K., Lee, K., Moran, C., Xie, Y., Honbo, D., Agrawal, A., Liao, W.K. and Choudhary, A., 2012. "User-Interest Based Community Extraction in Social Networks," in *The 6th SNA-KDD Workshop* (12), Beijing, China.

Papadopoulos, S., Kompatsiaris, Y., Vakali, A. and Spyridonos, P., 2012. "Community Detection in Social Media," *Data Mining and Knowledge Discovery*, (24:3), pp. 515-554.

Perrin, A. 2015. "Social Media Usage: 2005-2015." Pew Research Center, Washington, D.C. (October 8 2015), http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015, Retrieved January 30 2016.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. and Parisi, D., 2004. "Defining and Identifying Communities in Networks," in *Proceedings of the National Academy of Sciences of the United States of America*, (101:9), pp.2658-2663.

Ren, Y., Kraut, R. and Kiesler, S., 2007. "Applying Common Identity and Bond Theory to Design of Online Communities," *Organization studies*, (28:3), pp. 377-408.

Schaeffer, S.E., 2007. "Graph Clustering," *Computer Science Review*, (1:1), pp.27-64.

Shi, J. and Malik, J., 2000. "Normalized Cuts and Image Segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* (22:8), pp.888-905.

Xia, Z. and Bu, Z., 2012. "Community Detection Based on a Semantic Network," *Knowledge-Based Systems*, (26*)*, pp.30-39.

Yang, J. and Leskovec, J., 2013. "Overlapping Community Detection at Scale: a Nonnegative Matrix Factorization Approach," in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining,* Rome, Italy, pp. 587-596.

Yang, J. and Leskovec, J., 2015. "Defining and Evaluating Network Communities Based on Ground-Truth," *Knowledge and Information Systems*, (42:1), pp.181-213.

Yang, J., McAuley, J. and Leskovec, J., 2013. "Community Detection in Networks with Node Attributes," *13th IEEE International Conference on Data Mining (ICDM),* Dallas, TX, pp. 1151-1156.

Zhao, Z., Feng, S., Wang, Q., Huang, J.Z., Williams, G.J. and Fan, J., 2012. "Topic Oriented Community Detection through Social Objects and Link Analysis in Social Networks," *Knowledge-Based Systems*, (26*)*, pp. 164-173.