An Empirical Investigation of Like-Mindedness of Topically Related Social Communities on Microblogging Platforms

Kuntal Dey, Sivaji Bandyopadhyay
IBM Research Lab India, Jadavpur University
kuntadey@in.ibm.com, sbandyopadhyay@cse.jdvu.ac.in

Abstract

Microblog portals, like Twitter, are prominent online social networking platforms today. Users of such platforms exchange discussions and opinions around events and interests. Streaming algorithms have emerged to detect events early in their lifecycle from massive-scale microblogs. Methods have been developed to identify topical discussions evolving over time, on microblogs like Twitter, comprising of unstructured data and no explicit discussion thread. Such topical discussions comprise of semantically connected event clusters. However, no study on the characteristics of community structures formed around the topical discussion clusters have been proposed in literature. In the current work, we identify the structural communities, formed by the social network participants, for the topically connected semantic clusters, using modularity maximization algorithms. We observe better communities with significantly higher modularities within topical (semantic) clusters compared to global, showing like-mindedness of individuals participating in such communities. Our work is useful for applications that aim to leverage the pattern and structure of social spread of information.

1 Introduction

Social media has emerged as one of the largest pools of user-generated content. A plethora of social networking platforms have established over the better half of the past decade, thereby providing a level-playing platform for individuals to participate and interact in a social manner. Among the multitude of online social media platforms, the significant ones today comprise of social activity and interaction networks such as Facebook, microblogging networks such as Twitter, photo and image sharing and pinning platforms such as Pinterest and video sharing and distribution portals such as Youtube.

Of late, there has been a surge of research interest in propagation of topics, influence and interest over social network and media, including over social microblogging platforms like Twitter. A recent study by (Kwak, 2010) has established that Twitter helps topical information diffuse in a manner akin to news media. Twitter and other social networks, with their millions of users and connections along with multiple millions of text messages, are expected to have high levels of entropy by default. However, contrary to such expectation, recent work by (Narang, 2013) can identify social discussion threads on Twitter and other microblogging platforms, by socio-temporally correlating topically related text clusters.

It is of interest to note that unstructured microblogs like Twitter are hotbeds of conceptions of trending events. A typical characteristic of such trending events is that they tend to get involve individuals that are interested in those events. Such behavior apparently gets driven by similarity of interest of the participating individuals towards the topics associated with the event. However, the finding of (Narang, 2013) that topical discussions on microblogs, formed around events, show a tendency to evolve socially is interesting. This essentially introduces the significance of social familiarity in the propagation of information and interest along online social media. In addition, this helps distinguishing such evolving conversations around well-formed topics, from isolated expressions of topical interest.

While the finding by (Narang, 2013) is interesting, the study does not address any investigation pertaining to the characteristic of social communities participating in topical discussions. Clearly,

it is important to understand the true social nature of the communities formed by the participants in semantically connected discussion topics. Further, an attempt to conduct such a study will beg another question of research merit: are the communities formed by virtue of online discussions more like-minded compared to the globally formed structural community?

In this study, we propose addressing the above questions. To this, we propose using a well-known method, namely modularity, to measure the goodness of the communities formed around each semantically connected topical cluster. Using modularity, we also evaluate the goodness of inherent structural community formed in the social network graph. We compare the modularity values of the communities formed around the topical clusters and the global values. We demonstrate significantly higher modularity values for the topical communities, thereby underscoring the impact of like-mindedness, in the process of information flow, on top of the apparent entropy of microblogs. To the best of our knowledge, ours is the first study of its kind.

We use three large scale real-life Twitter datasets, namely Libya 2011 political turmoil, Egypt 2011 political turmoil and London 2012 Olympics, having thousands of users and up to millions of tweets, to conduct experiments. For all datasets, we observe significant presence of communities within topic clusters. We find significant improvement in modularity for communities formed within topic-clusters, compared to the global social graph based community structure. We believe this insight to be both novel and interesting. This leads to the observation that like-minded people with similarity in interest of topics are structurally better connected.

In summary, the main contributions of our work are the following.

- We investigate the goodness of the communities formed around topical discussions, using well-known measurement techniques.
- We provide a comparative evaluation the goodness of the topic-based communities against the global community structure formed in the topic-independent social network graph.
- We empirically demonstrate a trend of higher degree of like-mindedness of individuals participating in topical communities, compared to the like-mindedness of the overall graph.
- We demonstrate our findings on microblogging data using three real events.

The rest of our work is as follows. In Section 2, we explore related literature and discuss the state of the art. Subsequently, in Section 3, we discuss the problem settings in more detail, and provide an outline of our approach to solve the problem. We present our experimental results in Section 4. Finally, we conclude in Section 5.

2 Related work

Significant research has been conducted on content analysis of information discussed on social media sites (Kwak, 2010). (Grinev, 2009) demonstrate TweetSieve, a system that obtains news on any given subject by sifting through the Twitter stream. Along similar lines, Twinner by (Abrol, 2010) identify news content of a query by taking into account the geographic location and the time of query. (Nagar, 2012) demonstrate how content flow occurs during natural disasters.

Several ways to cluster social content have been studied. There has been work on clustering based on links between the users by doing agglomerative clustering, min-cut based graph partitioning, centrality based and Clique percolation methods (Porter, 2009), (Fortunato, 2007). Other approaches consider only the semantic content of the social interactions for the clustering (Zhou, 2006). More recently there has been work on combining both the links and the content for doing the clustering (Pathak, 2008), (Sachan, 2012). In (Narang, 2013) relationships between clusters are determined based on semantic, linkage and temporal information. Studies such as (Xu, 2000) have shown how local context analysis can improve the effectiveness of information retrieval. In the current work, we aim to study the phenomenon of social connections, across clusters, in the topically similar local social neighborhood.

Social network based community identification and graph structure analysis have been areas of research interest among academicians, and a significant number of prior studies have been conducted in these areas (Fortunato, 2007). Communities are not only useful in different kinds of academic studies, but also for practical applications (Jaho, 2011) (Modani, 2012).

Multiple schools of thoughts have emerged in the process of defining and identifying communities. One body of work is dedicated to identifying structural communities from input graphs, some of which are hard to find. This includes discovering cliques, quasi-cliques, k-cores, k-cliques, k-clubs and k-plexes (Dourisboure, 2007) (Gibson, 1998) (Hanneman, 2005) (Modani, 2008). Overlapping communities have been studied in (Chen, 2010) (Palla, 2005) (Sun, 2010).

Another body of work concentrates around partitioning the graph into subsets of vertices, such that, the connections of pairs of vertices within the subset of vertices are dense, while the connections of pairs of vertices across two subsets of vertices are of lesser density. This approach was proposed by (Girvan, 2002) and was taken further forward by (Clauset, 2004) and (Newman, 2004) in subsequent works. This was further followed up by spectral analysis techniques (Newman, 2006). The objective of these approaches are to maximize the modularity of partitioning a given input graph by identifying vertex subsets appropriately.

The BGLL algorithm by (Blondel, 2008) provides the fastest known algorithm for community finding based upon such graph partitioning. BGLL is known to provide communities with the highest modularity values among the known, state-of-the-art techniques. In our work, we use the BGLL algorithm to find the best possible graph partitioning based communities, with maximized modularity, in the fastest possible manner.

While our literature survey shows significant volumes of prior background art, which will be useful for us to solve the current problem, we do not find any work that attempts to solve our problem at hand. This establishes the novelty of our current work. At the same time, a review of the existing literature also amply motivates the need and timeliness for a study such as ours.

3 Problem settings and our approach

In this section, we describe the problem settings and propose the solution approach that we follow.

3.1 Problem settings

Objective: The objective of our work is to identify structural communities within topical clusters formed around semantic concepts, and derive insights about the characteristics of the communities. We aim to study the quality of communities thus formed, and compare these with the non-semantic communities that emerges from the topic-independent social network connection graph.

In order to meet our objective, there are a number of technical challenges to be overcome. We list the set of challenges below.

- First of all, we need to create the topic-based semantic clusters for a given event. This set of clusters will identify the independent set of discussions happening around the event under investigation.

- We now need to apply appropriate community detection algorithms in order to identify the communities formed by the social networks of the individuals participating in the topic clusters.

- Also, we find it interesting to report some of the basic characteristics of the topical communities discovered in the process.

- Finally, we need to be able to measure the goodness of these topical communities, as well the global communities, so that we compare across these communities and affirm the goodness of topical communities compared to non-topical ones.

3.2 Our approach

In this subsection, we outline our approach to solving the problem at hand. We first use an online streaming event detection algorithm for finding clusters from tweets (Weng, 2011). Following that, we attempt to discover graph partitioning based communities, formed by the set of the participants participating in each of the topical clusters found above. To this, we use a well-known modularity maximization algorithm - namely, BGLL (Blondel, 2008) - that creates disjoint graph communities with modularities being the highest in the known state of the art. We subsequently identify the communities on the overall social network graph also using BGLL, and thereby measure the global graph modularity. We investigate the modularity values of the topical clusters and the global graph modularity, in order to determine the comparative goodness of the communities found by each of the processes.

Basic notations and definitions

$\mathcal{E}$ denotes the list of events extracted from Twitter Stream. Since, event extraction is not the key focus of this work, we use a well-known online clustering algorithm ((Weng, 2011)) to generate event topics from streaming Tweet data, thereby creating clusters of tweets forming event topics. An event E^i is represented as $\{(K_1^i, K_2^i, \ldots, K_n^i), (L_1^i, L_2^i, \ldots, L_m^i), [T_s^i, T_e^i]\}$, where K^i denotes the set of keywords extracted from the tweets which form the event E^i, L^i is set of locations in event cluster and T^i is time period of the event. We use existing established methods for computing K, L and T. K contains *idf* vector and proper nouns (extracted by PoS tagging) from the tweets. L is generated by using Standford's NLP Toolkit and associated Named Entity Recognizer. T is simply the time of first and last tweet in the event cluster.

Forming topical clusters

As mentioned previously, we use the online streaming event detection algorithm by (Weng, 2011) for finding topical (semantic) clusters from tweets. More specifically, we use the *EDCoW* method to form topical clusters. This method uses *cross-correlation* methods of (Orfanidis, 1996) in order to measure similarity between two signals. The intuition behind this event detection algorithm, in order to form topical clusters, is to compute similarity between words, and thereby group sets of words with similar burst patterns.

In signal processing, the cross-correlation between two signals, represented as functions $f(t)$ and $g(t)$, is given by

$$(f * g)(t) = \sum f^*(\tau) g(t + \tau) \qquad (1)$$

In Equation 1, the term f^* denotes the complex conjugate of f. Cross-correlation computation is used to shift one signal (in this case, g in Equation 1), and calculates the dot product between the two signals. In other words, it measures the similarity between the two signals as a function of a time-lag applied to one of them.

Computation of cross-correlation being a pairwise operation, it is expensive to measure the cross-correlation for all signal pairs for any massively populated microblog like Twitter. However, as pointed out by (Weng, 2011), many of these signals happen to be trivial, and this has been proved by empirical investigation of a large dataset. So while the algorithm for finding out cross-correlation is $\mathcal{O}(n^2)$, the authors go on to show that practically, after applying filters for eliminating undesirable outliers and making median-based selections, less that 5% of the total words tend to remain. Therefore, the quadratic complexity of the process remains tractable for practical purposes. After this, it applies Newman's modularity techniques (Newman, 2004) (Newman, 2006) for graph partitioning.

Further, the *EDCoW* algorithm by (Weng, 2011) quantifies event significance, as each microblog post (like a Tweet) is associated with only a few words because of their short length, and the algorithms requires at least 2 words to be a part of the message for functioning. One can denote the subgraph corresponding to an event c as: $G_c = (V^c, E^c, W^c)$. In this, V^c is the vertex set and $E^c = V^c \times V^c$. W^c contains the weights of the edges that are given by a portion of the correlation matrix $\mathcal{M}$. The event *significance* is hence defined by (Weng, 2011) as:

$$\epsilon = (\sum w_{ij}^c) \times \frac{e^{1.5n}}{(2n)!}, n = |V^c| \qquad (2)$$

Equation 2 comprises of two parts. The first part sums up all the cross correlation values between signals associated with an event. The second part discounts the significance if the event is associated with too many words. The higher is the ϵ, the more significant is the event. Also, the *EDCoW* algorithm by (Weng, 2011) filters events with exceptionally low value of ϵ, such as ($\epsilon \ll 0.1$).

Community finding and modularity

We now attempt to find social communities formed by the participants of each topical (semantic) cluster, by investigating each of the clusters. Let the set of clusters, $\{C\}$, comprise of k clusters, namely $\{c_1, c_2, ...c_k\}$. For our solution, for any cluster $c_i \in \{C\}$, we first consider the set of vertices, V_{c_i}. We then construct the induced subgraph of the set of vertices, based upon social connections, constructing edges E_{c_i} such that the pair of vertices constituting the edge belong to the same cluster. This process leads to the construction of an induced subgraph $G_{c_i} = \{V_{c_i}, E_{c_i}\}$. Therefore, at the end of this process, for the k clusters, we have a set of k induced subgraphs, namely $G_{c_1}, G_{c_2}, ..., G_{c_k}$.

We now attempt to find graph partitioning based communities on each of the induced subgraphs. For this, we choose a graph partitioning based algorithm, which is a variant of the modularity maximization algorithm family. One of the most noteworthy characteristics of graph partitioning based approaches is that the communities derived are necessarily non-overlapping in nature. In other words, there will be no single vertex, at any time, belonging to one community, and also belong to another community at the same time.

Modularity was proposed by Newman in (Newman, 2004), and was enhanced with spectral methods in (Newman, 2006). Modularity is a quantity that attempts to measure the difference of, the actual sum of weight of edges that lie within a given component after the graph is partitioned, and the expected sum of edge weights if the edges were drawn at random by sheer probability. Higher modularity values indicate better partitioning of the graph such that communities of better quality get grouped together.

Let A_{ij} be the number of the edge between vertices i and j. Here, A_{ij} is an element of the adjacency matrix of the social network. One can easily show that for a network comprising of m edges, the expected number of edges connecting vertices i and j, if the positions of the edges are randomized, is given by $k_i k_j/2m$, where k_i and k_j are the degrees of node i and j respectively. Hence, the actual number of edges between i and j minus the expected number of edges is $A_{ij} - k_i k_j/2m$. The modularity Q is derived by adding all the pairs of vertices belonging to the same community. If we label the communities and define s_i to be the label of the community to which node vertex i belongs, then we get modularity as:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta_{s_i, s_j} \qquad (3)$$

In the above, δ_{s_i,s_j} is the Kronecker delta, and s_i and s_j are the index of the subgraph that the vertices v_i and v_j belong to respectively. The leading constant $\frac{1}{2m}$ is included by convention: it normalizes Q to measure fractions of edges rather than total numbers but its presence has no effect on the position of the modularity maximum. The objective here is to partition the input graph G such that the value of the modularity, namely Q, is maximized.

In case of a weighted graph like ours, one can replace A_{ij} by the weighted edge w_{ij}, thereby leading to the following equation:

$$Q = \frac{1}{2m} \sum_{ij} (w_{ij} - \frac{k_i k_j}{2m}) \delta_{s_i, s_j} \qquad (4)$$

Newman's spectral graph based approach to optimize modularity (Newman, 2006) leverages the above concept. It initially creates a modularity matrix B on a given graph G, in which the elements can be obtained by:

$$B_{ij} = w_{ij} - \frac{d_i . d_j}{2m} \qquad (5)$$

Following this, the symmetric matrix B undergoes an eigen-analysis process, by which the largest eigenvalue and the corresponding eigenvector $\vec{v}$ are found. Finally, G is split into two subgraphs based on element signs in $\vec{v}$. The spectral method is recursively applied to each of the two subgraphs, thereby dividing them into smaller and smaller subgraphs, as long as the overall modularity of the graph partition keeps increasing, thereby forming communities with high modularity.

The BGLL algorithm

After introducing modularity, we now focus on a specific modularity maximization algorithm that we use in the current work, namely the BGLL algorithm. This is the technique proposed by (Blondel, 2008), and till date is known to be one of the fastest algorithms that also partition the graph to obtain the highest modularity values in the known state of the art. We use BGLL to discover modularity-maximized communities in each of the topic-cluster-induced subgraphs, as well as the global graph.

The BGLL algorithm has two phases. In the first phase, each node is assigned to a singleton cluster. The clusters are now reorganized by moving a vertex into the group of neighborhood vertices, and thereby measuring the change of modularity. The vertex is retained in the group for which the modularity gain is positive and maximum. This process is repeated for all vertices, until no further modularity value improvement remains possible.

Here, the modularity gain, ΔQ, is computed as:

$$\begin{aligned} \Delta Q = [&\frac{\sum_{in} + 2k_{i,in}}{2m} - (\frac{\sum_{tot} + k_i}{2m})^2] \\ -[&\frac{\sum_{in}}{2m} - (\frac{\sum_{tot}}{2m})^2 - (\frac{k_i}{2m})^2] \end{aligned} \qquad (6)$$

Here, $\sum_{in}$ is the sum of edge weights inside a given community, $\sum_{tot}$ is the sum of the weight of incoming edges to the community, k_i is the sum of the weights of the edges incident to vertices within the community, $k_{i,in}$ is the sum of the weights of the edges from i to vertices in the community, and m is the sum of the weights of all the edges in the social network.

The second phase aims to build a network by treating the communities found in first phase as vertices of a graph, thereby creating hypernodes, and creating an edge between two hypernodes where the weight is given by the sum of weights of the edges in the communities.

Goodness of our method

After having found the communities for each of the topical (semantic) clusters as well as for the overall social network graph, we now attempt to measure the goodness of our findings. To this, we observe the distribution of the modularity values of the induced subgraphs thus formed. We further measure the various statistical parameters of the modularity distribution.

The number of clusters where $\mathcal{M}_i > \mathcal{M}$, that is, the number of clusters where modularity of c_i is higher than the global graph modularity, is an interesting measure. Another interesting measure is the mean cluster modularity, $\mu(\mathcal{M})$, and a comparison of this value with the global graph modularity. These values show the goodness of our findings. As we shall observe in the subsequent section, our experiments indicate encouraging values of modularity and its distribution.

Another interesting observation is that, by virtue of our construction methodology, each of the communities we look at, will completely belong to one topical (semantic) cluster, and never span across cluster boundaries. In essence, these communities can be viewed as social communities formed on graphs of shared interest - namely, the topic central to the corresponding cluster. Therefore, the higher the value of modularity, the higher is the degree of life-mindedness among the cluster members.

Our experiments show that the social network graphs tend to exhibit a higher degree of like-mindedness compared to the global graph, as inferred from the modularity value distribution over the set of induced subgraphs.

4 Experimental results

We now proceed on to conduct the experiments, following the algorithms and sequence described in Section 3.

4.1 Data description

The first step is data collection. We collect Twitter data from three large-scale events: (1) A 2011 Libya political turmoil that had created significant impact on social media, (2) A Egypt 2011 political turmoil data that also had a significant footprint on Twitter and (3) The London Olympics 2012 data. The data was collected using (Libya OR Gaddafi), (Egypt OR Protest) and (Olympic OR Olympics) as target keywords respectively. This implies, all the tweets used for experimentation contain at least one of the two above-mentioned keywords, or both, for each corresponding dataset.

We further collect the social network (*followers*) data for these users. We use these datasets to qualitatively inspect the goodness of our extended semantic edge generation algorithm. Table 1 shows the statistics of all the 3 datasets used in our study.

4.2 Forming the baseline graphs

Since event cluster detection is not the focus of our work, we have used the online clustering algorithm by (Weng, 2011) to generate the event clusters from the given tweets, as described in Section 3. The outcome of applying this algorithm is the set of clusters, as illustrated in Table 1.

For each cluster discovered in the process, we now require constructing the induced subgraph of the participants of the cluster with at least one tweet, over the social network followership edges. In order to construct this, we retain all the edges, in which, both the endpoints belong to the same cluster. At the same time, we discard any cross-cluster edge. Please note that because of the method of our construction, an edge can potentially belong to multiple clusters - in other words, more than one cluster could have overlap in terms of participants and edges; however, there can be no edge that would not have both its endpoints in the same cluster for a given graph partition.

After construction of the semantic (topical) clusters and the social graphs, we attempt to identify the communities within each of the clusters, and measure the modularities of these communities, in order to meet our stated objective.

Table 1: The columns in the table show a) keywords used to search Twitter to collect the dataset, b) dates for which the data was collected, c) number of tweets collected, d) number of clusters and e) number of users on the social network

Dataset	Keywords	Timespan	Tweets	Clusters	Number of Users
Egypt	Egypt, Protest	1 - 4 Mar'11	60,948	142	37,961
Olympics	Olympics, Olympic	27 Jun - 13 Aug'12	2,319,519	299	1,313,578
Libya	Libya, Gaddafi	4 - 24 Mar'11	1,011,716	1,344	83,177

Table 2: Comparison of the modularity of the global graphs and those of the topical (semantic) cluster graphs. Please note that *% Topical Clusters Mod* $>$ *Global* denotes the percentage of topical clusters in each dataset in which the modularity values are higher than the global modularity (zero values excluded).

Dataset	Number of Clusters	Global Modularity	% Topical Clusters Mod $>$ Global	Mean Topical Cluster Modularity	Maximum Modularity
Egypt	142	0.48	75.18%	0.59	0.94
Olympics	299	0.56	78.64%	0.71	0.98
Libya	1,344	0.66	12.1%	0.40	0.83

4.3 Experimental results

After forming the baseline graphs and discovering the topical (Semantic) clusters, we now conduct the following actions.

- We identify the communities within each cluster, and subsequently find the modularity of each cluster based upon the communities formed.
- We identify the communities in the original input graph, and find the modularity of this community distribution.
- We compare the two modularities found by the above processes.

Table 2 shows basic statistics of the modularities we find for the individual clusters, as well as the modularity of the overall graph. As clearly seen from the table, for all the datasets, the modularity values are significant. Figure 1 visually represents the distribution of the modularity values, in form of a scatter plot.

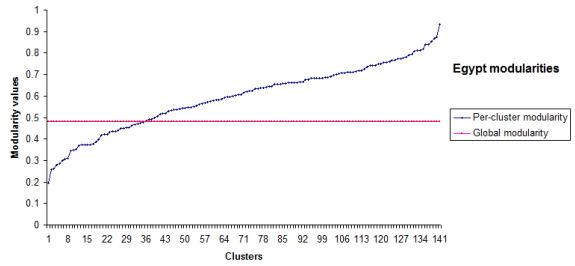
From our experimental observations, it is obvious that in case Egypt and Olympics, a significant fraction (more than $3/4^{th}$ in both the cases) of the modularity values at the per-cluster level is higher than the global modularity. In case of Libya, this fraction is much smaller (only $1/8^{th} of the dataset$), but still there is evidence of this phenomenon. Figure 1(a) and Figure 1(b) therefore visually show most of the topical cluster distribution over the global level, thus showing pronounced impacts of presence of like-mindedness. Figure 1(c) shows the later parts (right-hand side) of the modularity distribution cross the global level though the values are below the global level for most of the earlier parts (left-hand side) for this graph, showing presence, albeit somewhat less pronounced impacts, of like-mindedness.

In fact, the maximum modularities in some cases are higher than 0.9, the highest being 0.98 for Olympic dataset, which is surprisingly high, indicating a near-perfect community structure. The mean modularity of topical clusters, the maximum modularity of topical clusters and the percentage of clusters that have a modularity value higher than the corresponding global modularity - all clearly indicate the goodness of our approach.

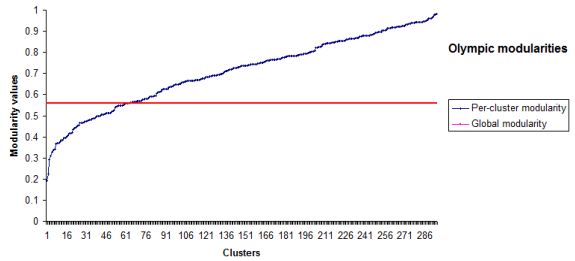
4.4 Discussions

As obvious from the experimental results, we observe high modularity values (and associated measurements) for topical communities formed around semantic clusters, compared to the global modularity. These higher modularity values and the associated factors measured, indicate the goodness (formation of comparatively stronger communities) of the topic-based communities, compared to the global community structure formed in the topic-independent social network graph.

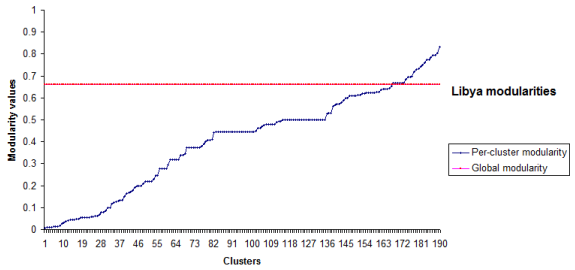
The semantic clusters are formed based upon discussion topic similarity. Therefore, the signif-



(a) Egypt modularity values



(b) Olympic modularity values



(c) Libya modularity values

Figure 1: Modularity value distribution for the three datasets, for topical (semantic) clusters and the global graph

icantly high presence of stronger communities indicate topical like-mindedness of the community participants in the communities within semantic clusters.

This helps us to draw the conclusion that on microblogging platforms like Twitter, there is a significant impact of topical like-mindedness on formation of social communities.

5 Conclusions

In this work, we studied the formation and characteristics of social communities within semantically related topical clusters. We used a well-known technique to identify the semantically related topical clusters. Subsequently, we used another well-understood method to discover communities within each of the topical clusters. We found the modularities of each of the communities formed along the topical clusters. We proved that these modularities are significantly higher than that of the communities formed from the underlying structural graph.

Our work leads to the observation that like-minded people with similarity in interest of topics are structurally better connected. This also proves that the spread of information is likely to be social, and more along like-minded individuals forming a community based upon their familiarity network.

This work will be useful for social networking and other families of applications that aim to leverage the pattern and structure of social spread of information. As future work, we propose to identify implicit interest groups from the nature of information propagation. We also believe that identifying contextually related topics based upon the underlying social network will be a novel and interesting study as a next step.

References

Abrol S., Khan L.: *Twinner: Understanding news queries with geo-content using twitter.* In: Proceedings of the GIS (2010).

Allen J. F.: *Maintaining Knowledge about Temporal Intervals.* In: Communications of the ACM (1983).

Blondel V.D., Guillaume J.L., Lambiotte R., Lefebvre E.: *Fast unfolding of communities in large networks.* In: J. Stat. Mech. P10008 (2008).

Chen W., Liu Z., Sun X., Wang Y.: *A game-theoretic framework to identify overlapping communities in social networks.* In: Data Min. Knowl. Discov., 21(2):224–240 (2010).

Clauset A., Newman M. E. J., Moore C.: *Finding community structure in very large networks.* In: Phys. Rev. E, 70(066111) (2004).

Coombs C. H., Dawes R. M., Tversky A.: *Mathematical psychology: An elementary introduction* In: Englewood Cliffs, NJ: Prentice-Hall (1970).

Dourisboure Y., Geraci F., Pellegrini M.: *Extraction and classification of dense communities in the web.* In: WWW, pages 461–470 (2007).

Fortunato S., Barthelemy M.: *Resolution limit in community detection.* In: Proceedings of the National Academy of Sciences, 104(1):36-41 (2007).

Gibson D., Kleinberg J., Raghavan P.: *Inferring web communities from link topology.* In: HYPERTEXT, pages 225–234 (1998).

Girvan M, Newman M. E. J.: *Community structure in social and biological networks.* In: Proc. Ntl. Acad. Sci, USA, 99(7821) (2002).

Grinev M., Grineva M., Boldakov A., Novak L., Syssoev A., Lizorkin D.: *Tweetsieve: Sifting microblogging stream for events of user interest.* In: Proceedings of the SIGIR (2009).

Hanneman R. A., Riddle M.: *Introduction to Social Network Methods.* In: University of California, Riverside, CA (2005).

Jaho E., Karaliopoulos M., Stavrakakis I.: *Iscode: a framework for interest similarity-based community detection in social networks.* In: *INFOCOM WORKSHOPS*, pages 912–917. IEEE (2011).

Kwak H., Lee C., Park H., Moon S.: *What is Twitter, a Social Media or a News Media.* In: Proceedings of the WWW (2010).

Lin D.: *An information-theoretic definition of similarity.* In: Proceedings of the International Conference on Machine Learning (1998).

Modani N., Dey K.: *Large maximal cliques enumeration in sparse graphs.* In: CIKM, pages 1377–1378 (2008).

Modani N., Gupta R., Nagar S., Sanigrahi S., Goyal S., Dey, K.: *Like-minded communities: bringing the familiarity and similarity together.* In: WISE (2012).

Nagar S., Seth A., Joshi A.: *Characterization of Social Media Response to Natural Disasters.* In: Proceedings of the WWW (2012).

Narang K., Nagar S., Mehta S., Subramaniam L.V., Dey K.: *Discovery and analysis of evolving topical social discussions on unstructured microblogs.* In: European Conference on Information Retrieval (2013).

Newman M. E. J.: *Fast algorithm for detecting community structure in networks.* In: Phys. Rev. E, 69(066133) (2004).

Newman M. E. J.: *Modularity and community structure in networks.* In: Proc. Natl. Acad. Sci., 103:85778582 (2006).

Orfanidis S. J: *Optimum Signal Processing.* In: McGraw-Hill (1996).

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

Pathak N., DeLong C., Banerjee A., Erickson K.: *Social topics models for community extraction.* In: Proceedings of the 2nd SNA-KDD Workshop (2008).

Porter M. A., Onnela J. P., Mucha P. J.: *Communities in networks.* In: Notices of the American Mathematical Society, 56(9), pp. 1082-1097 (2009).

Sachan M., Contractor D., Faruquie T. A., Subramaniam L. V.: *Using content and interactions for discovering communities in social networks.* In: Proceedings of the WWW (2012).

Sun H., Huang J., Han J., Deng H., Sun Y.: *SHRINK: A Structural Clustering Algorithm for Detecting Hierarchical Communities in Networks*. In: CIKM (2010).

Weng J., Lee B.S: *Event detection in Twitter.* In: ICSWM - Proceedings of the AAAI conference on weblogs and social media (2011).

Xu J., Croft. W. B.: *Improving the effectiveness of information retrieval with local context analysis.* In: ACM Trans. Inf. Syst., 18(1):79-112 (2000).

Zhou D., Manavoglu E., Li J., Giles C. L., Zha H.: *Probabilistic models for discovering e-communities.* In: Proceedings of the WWW (2006).