# Identifying Focal Patterns in Social Networks

Fatih Sen, Rolf T. Wigand, Nitin Agarwal, Debanjan Mahata and Halil Bisgin
Department of Information Science,
University of Arkansas at Little Rock, U.S.A
{fxsen, rtwigand, nxagarwal, dxmahata, hxbisgin}@ualr.edu

***Abstract*—Identifying authoritative individuals is a well-known approach in extracting actionable knowledge, known as "Knowledge Representation", in a social network. Previous researches suggest measures to identify influential individuals, however, such individuals might not represent the appropriate context (relationships, interactions, etc.). For example, it is nearly an impossible task for a single individual to organize a mass protest of the scale of Occupy Wall Street. Similarly, other events such as the Arab Spring, coordinating crisis responses for natural disasters (e.g., the Haiti earthquake), or even organizing flash mobs would require a key set of individuals rather than a single or the most authoritative one. These events demonstrate the need and importance of examining *influential structures* rather than single individuals in social networks. A new methodology is proposed to identify such influential structures and recognizing their importance. The proposed methodology is evaluated empirically with real-world data from NIST's Tweets2011 corpus. We also introduce a novel and objective evaluation strategy to ascertain the efficacy of the focal patterns. Challenges with future research directions are outlined.**

***Keywords*-social media, focal patterns, event analysis**

## I. INTRODUCTION

The modern world has already witnessed how social media became a key player in organizing and coordinating events, such as 'The Arab Spring', 'Occupy Wall Street', 'Flash Mobs', 'The London Riots', and many other real-life events. Such events could be successfully co-ordinated, only by the collective efforts and influence of groups of enthusiastic people instead of an individual. Considering an individual person as a node in a social network, the authoritative approach identifies the most centralized nodes with neighbors. However, a node on its own may not represent the context. Therefore, there is a need to identify and represent the central authoritative structures besides the authoritative nodes, which is the major aim of this paper.

Community based analysis deals with the detection of communities or clusters, and, might overlook the small-focused structures. The need of identifying such hidden structures in terms of context and specific interaction patterns, and to study their importance, motivated us for the presented work. Therefore, unlike other clustering and central algorithms, focal patterns are identified instead of communities and authoritative nodes in a network, as shown in Fig. 1. The main purpose of this research is to detect the focal patterns in various networks and to represent knowledge in terms of the detected focal patterns.

The primary contributions of the research are:

1) Defining focal patterns (Section III).
2) Proposing a methodology to extract focal patterns (Section IV).
3) Demonstrating the usefulness of detecting focal patterns over influential nodes on a sample Twitter network of users tweeting on 'The Egyptian Revolution' of 2011 (Section V).

## II. RELATED WORK

Finding influential nodes in a social network has been studied in the context of webpage ranking, especially the algorithms such as PageRank [4] and HITS [6]. PageRank would assign a numerical weight for each blog post to "measure" its relative importance [5]. Researchers have also studied social networks from the perspective of information diffusion and have identified key players who maximize its spread [7]. Gruhl et al. [8] studied information diffusion of various topics in a social network, drawing on the theory of infectious diseases. A general cascade model [9] was also adopted. An interesting problem related to viral marketing [10], [11], [12] is how to maximize the total influence in the social network by selecting a fixed number of nodes in the network. A greedy approach can be adopted to select the most influential node in each iteration after removing the selected nodes. This greedy approach outperforms PageRank, HITS and ranking by number of citations, and is robust in filtering splogs (spam blogs) [13]. Leskovec et al. [14] proposed a submodularity based approach to identify the most important nodes which outperforms the greedy approach. However, the proposed method in this paper, is different from all such works mentioned above, and finds influential structures rather than influential nodes from a given network.

## III. PROBLEM DEFINITION

In the real-world, communication structures include interaction patterns leading to questions on which patterns are most influential in the whole network. It is impossible to determine these influential patterns with one single node, or authoritative node. Also, an influential interaction pattern is different from a community, because the patterns indicate the backbone of the network, whereas a community may not. This is essential not only for the network of communication structures but also for organizational networks or for any other social network in which, the interaction patterns of densely connected actors are involved. Therefore, we are inquisitive to find the most influential interaction structures instead of influential nodes.

978-1-4673-4794-5/12/$31.00 

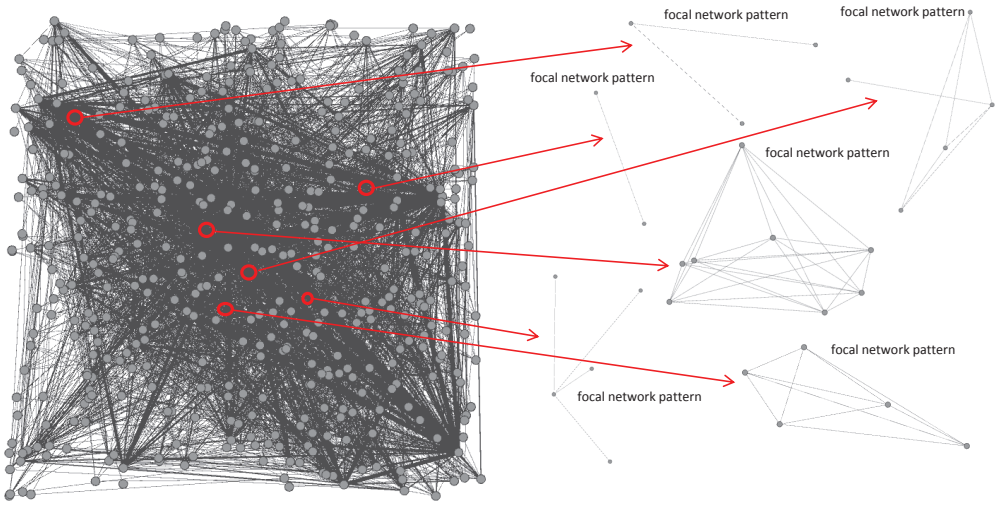



Fig. 1. The identification of desired focal network patterns in a social graph of 442 vertices and 3171 edges. The network data has been collected from http://www.livejournal.com, which is a social media platform where users share common passions and interests. The ultimate goal is to extract the hidden structures, which are densely connected with two, three, four, five, etc. nodes. Please note that a pattern is not restricted to contain a certain number of vertices. For example, in this figure, focal network patterns range from two to eight vertices.

In simpler terms, an influential interaction pattern is called a focal pattern for the network. A network can have more than one focal pattern with varying degrees of authoritativeness. Further, a focal pattern should have at least two nodes. Given a graph G with vertices V and edges E. We need to identify the most authoritative and smallest set of nodes forming the focal patterns F which are embedded in G. Each focal pattern is unique so that a focal pattern cannot be a subset of any other focal pattern. Formally, a focal structure can be defined as follows:

*Given a social network G = (V,E), where V is the set of vertices and E is the set of edges. Focal patterns in G are defined by* $F = \{G'\}$, *where* $G' = (V', E')$ *and* $V' \subseteq V$ *and* $E' \subseteq E$. *For all i and j,* $i \neq j$, $G_i \in F$ *and* $G_j \in F$, *such that no two focal patterns can subsume each other, or* $G_i \nsubseteq G_j$ *and* $G_j \nsubseteq G_i$. *Assuming we have an influential score for a focal pattern* $G_i$, $I(G_i)$. *F contains k most influential focal patterns* $G_{j_1}, G_{j_2}, ..., G_{j_k}$ *ordered according to their influential score such that* $I(G_{j_1}) \geqslant I(G_{j_2}) \geqslant ... \geqslant I(G_{j_k})$.

## IV. Proposed Methodology

A methodology, called f-patterns, is proposed in Section IV-A for identifying focal patterns for unweighted network graphs. Later, this approach is extended in section IV-B to consider weighted networks as well.

### A. The Proposed Approach: f-patterns for unweighted graphs

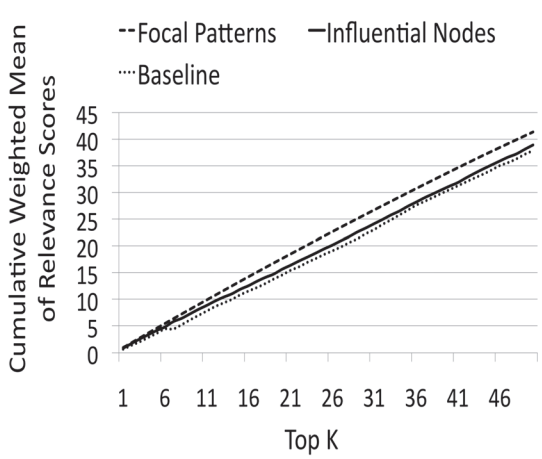
1) Apply the Louvain Method [2], [3] to the current(whole) graph and get the partitions along with the modularity value.
2) If the modularity value is zero (Q=0) return the graph (pattern).
3) Otherwise, get the sub-graphs and start iterating through all sub-graphs.
4) Consider each sub-graph as a whole graph and repeat the same steps (1, 2 and 3) recursively for this sub-graph.

The whole graph is partitioned using the modularity method [1] without any threshold values. The network patterns for the social graph is generated by applying the Louvain method. The modularity of a graph (or sub-graph) is zero (Q=0) if it cannot be partitioned anymore. Such a graph is the desired network pattern which is the main goal of this research. If a graph has sub-graphs, i.e. the modularity value is not zero (Q!=0), then each sub-graph is partitioned in a recursive manner until a modularity value of zero (a focal pattern) can be obtained.

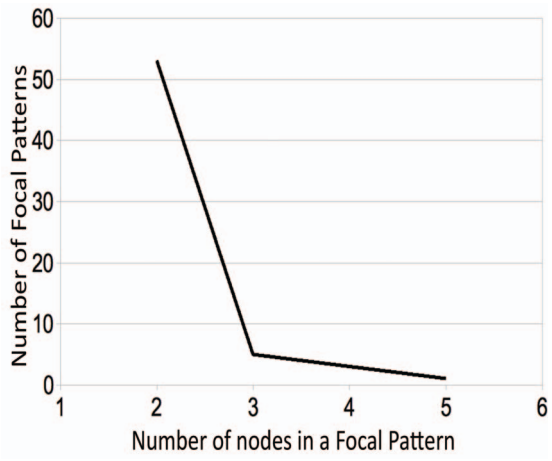
### B. The Proposed Approach: f-patterns for weighted graphs

1) Obtain the threshold values by getting the cumulative distribution of the number of the edges and edge weights.
2) Begin with the first threshold value.
3) Filter the graph with the threshold value.
4) Apply the Louvain method to the current graph and get the partitions along with the modularity value.
5) If the modularity value is zero (Q = 0) return the graph (network pattern).
6) Otherwise, get the sub-graphs and start iterating through all sub-graphs.
7) Consider each sub-graph as a whole graph and repeat the same steps (3, 4 and 5) recursively for this sub-graph.
8) Next threshold value.

The first part of the f-patterns approach (above) partitions the graph recursively and generates the desired network focal patterns. However, it does not solve the overlapping issue. The extended approach aims to solve this challenge by filtering the whole graph according to the edge weights, first, and then starting the partitioning process from the filtered




(a) Cumulative weighted mean of relevance scores of the named entities in the top K focal patterns and the influential nodes.




(b) Distribution of nodes in the detected focal patterns.

Fig. 2. Evaluation of proposed approach, f-patterns on Tweets2011 dataset provided by TREC.

graph. Basically, the graph is first filtered by various edge weights (threshold values) and the same process is applied as mentioned above. Recall that the filtering mechanism of this method leads to various focal and overlapping structures.


Fig. 3. Word cloud of the named entities related to Egyptian Revolution.

## V. RESULTS AND DISCUSSION

We performed our experiment on the Tweets2011[1] dataset provided by TREC[2]. The dataset provided identifiers for approximately 16 million tweets sampled between January 23rd and February 8th, 2011. This is also the time when the major events of 'The Egyptian Revolution' took place. These tweets were collected from Twitter by using the Twitter API. We further extracted 12000 tweets related to "Egyptian Revolution" and identified 3727 unique users. Twitter API was again used for getting the friends and the followers of each of these users. A network was constructed comprising of these users by analyzing their friends and followers lists. The tweets posted by each user were concatenated and assigned to each of them.

We also obtained 234 sources related to "Egyptian Revolution", 88 sources related to "Libyan Revolution" and 77 sources related to "Tunisian Revolution" from GlobalVoices[3]. GlobalVoices is a portal where bloggers and translators work together to make reports of various real-life events, from blogs and citizen media everywhere. These sources are manually curated and translated into English from their native language. The three events considered were the major events that took place in the middle-east parallely, and had gained wide popularity in the social media. We took three events in order to identify event-specific named entities that are very closely related to a particular event. We used the sources obtained from GlobalVoices to identify the named entities (name, place, organization, etc) which are highly relevant to "Egyptian Revolution". All the named entities were extracted using AlchemyAPI[4]. The popular Tf_iDf measure, was used for getting relevance scores of the named entities mentioned across the three events. Finally, a set of named entities were identified which were highly relevant to 'The Egyptian Revolution" of 2011. Some of them are shown in Fig 3. The proposed methodology resulted in 59 focal patterns with average size of 2.13 nodes per focal pattern in two levels of recursion.

In order to evaluate the focal patterns we came up with a novel strategy. We evaluated it in terms of the relevant information it presented and how quickly it helped us in learning valuable information about "The Egyptian Revolution" from the segregated tweets. Thereafter, we took the following steps:

1) Focal patterns and influential nodes were obtained from the network of Twitter users. We used the HITS algorithm to obtain the influential nodes. In addition, a baseline approach was devised. For each focal pattern, random nodes were selected from the pool of 3727 nodes. The number of nodes selected for each focal

[1]http://trec.nist.gov/data/tweets/
[2]http://trec.nist.gov/
[3]http://globalvoicesonline.org
[4]http://alchemyapi.com

pattern was equal to the number of nodes contained in the respective focal patterns. These combined units of nodes mapped with each focal pattern were taken as the baseline.

2) Named entities were extracted from the tweets of each user using AlchemyAPI. The weighted mean of the relevance scores of the named entities mentioned in these tweets were calculated, in order to find how relevant the tweets are with respect to the event. Weighted mean was used in order to take into account the frequency of occurrence of the named entities along with their relevance score, and to get a balanced score.
3) The influential nodes were arranged in a decreasing order of their influence score. Whereas, the focal network patterns were arranged in the decreasing order of the weighted mean of the relevance scores of the named entities mentioned in the tweets posted by the users in them.
4) The cumulative sum of the relevance scores (Y-axis) were obtained and plotted in a chart (Fig 2.) for the top K (X-axis, K=50 in our case) influential nodes, focal patterns and the baseline units. We took the cumulative sum in order to show the gain in relevant information about the event as we go on analyzing more and more focal patterns, influential nodes and the baseline units mapped with each focal pattern, respectively.

We can conclude from the chart that the focal patterns give more relevant information than obtained from the influential nodes as well as the randomly selected baseline units. They help us to quickly learn about the event by providing relevant information at a faster rate. Thus, the focal patterns obtained using our method shows promises as potential hubs of information related to the chosen event. Such an observation truly justifies the usefulness of focal patterns and shows its utility in analyzing events, and for obtaining event related valuable information from a social network website.

## VI. Future Work and Conclusion

Recognizing the importance of influential structures in social networks, the proposed research focuses on developing a framework for identifying such structures. The interactions within the nodes of these structures form the context, which would provide a deeper understanding of the above-mentioned phenomena.

In this work, we propose a methodology for identifying focal network structures in a given network. We also show how the focal structures help in finding valuable information from a social network, and how it performs better than the influential nodes in doing such a task. Such a result further demonstrates the need and importance of examining influential structures rather than single individuals in social networks.

The suggested framework can also be applicable to many areas, such as recommendation systems, semantic web, marketing, advertising, information diffusion studies, and search engine indexing, among others. Also, the research areas pertaining to network structural concepts; such as social capital, strength of weak ties, trust, diversity in network, bandwidth, similarity of nodes/structure/homophily, etc., can benefit from the proposed research. To the best of our knowledge, the proposed research is the first effort in identifying influential structures in a social network. Moreover, the methodological contributions of the research will help in analyzing and understanding real-world phenomena and advance foundational sociological concepts. We look forward to work further on the methodology and make it more robust and scalable in the near future.

## Acknowledgments

The research is supported in part by grants from the US Office of Naval Research (Award: N000141010091) and the US National Science Foundation (Awards: IIS-1110868, IIS-1110649).

## References

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

[2] J. M. Pujol, V. Erramilli, and P. Rodriguez. Divide and Conquer: Partitioning Online Social Networks, http://arxiv.org/abs/0905.4918v1, 2009.

[3] J. Haynes, and I. Perisic. (2009). Mapping Search Relevance to Social Networks. In: Proc. of 3rd Workshop on Social Network Mining and Analysis. *International Conference on Knowledge Discovery and Data Mining*, Paris, France.

[4] S. Brin and L. Page. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the Seventh International Conference on World Wide Web*, pp. 107-117, 1998. doi: 10.1016/S0169-7552(98)00110-X

[5] N. Agarwal, H. Liu, L. Tang, and S. Yu. Philip. (2008). "Identifying Influential Bloggers in a Community", *1st International Conference on Web Search and Data Mining (WSDM08)* , pp. 207-218. Stanford, California.

[6] J. Kleinberg. (1998). Authoritative Sources in a Hyperlinked Environment. In *9th ACM-SIAM Symposium on Discrete Algorithms*. doi: 10.1145/324133.324140

[7] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. (2007). Cascading Behavior in Large Blog Graphs. In *SIAM International Conference on Data Mining*.

[8] D. Gruhl, D. Liben-Nowell, R. Guha, and A. Tomkins. (2004). Information Diffusion Through Blogspace. *SIGKDD Exploration Newsletter*, 6(2):4352.

[9] J. Goldenberg, B. Libai, and E. Muller. (2001). Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters*, 12:211223.

[10] M. Richardson and P. Domingos. (2002). Mining Knowledge-Sharing Sites for Viral Marketing. *In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 6170, New York, NY. ACM Press.

[11] D. Kempe, J. Kleinberg, and E. Tardos. (2003). Maximizing the Spread of Influence Through a Social Network. *In Proceedings of the KDD*, pages 137-146, New York, NY, USA. ACM Press.

[12] W. Chen, Y. Wang, and S. Yang. (2009). Efficient Influence Maximization in Social Networks. *In KDD 09 Proceedings of the 15th ACM SIGKDD International Conference on Knowledge discovery and Data Mining*, pages 199208, New York, NY.

[13] A. Java, P. Kolari, T. Finin, and T. Oates. (2006). Modeling the Spread of Influence on the Blogosphere. *In Proceedings of the 15th International World Wide Web Conference*. University of Maryland, Baltimore County.

[14] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. (2007). Cost-Effective Outbreak Detection in Networks. *In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 420429, New York, NY.