Graph Summarization for Geo-correlated Trends Detection in Social Networks

Colin Biafore Faisal Nawab
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93106
colinbiafore@umail.ucsb.edu nawab@cs.ucsb.com

ABSTRACT

Trends detection in social networks is possible via a multitude of models with different characteristics. These models are pre-defined and rigid which creates the need to expose the social network graph to data scientists to introduce the human-element in trends detection. However, inspecting large social network graphs visually is tiresome. We tackle this problem by providing effective graph summarizations aimed at the application of geo-correlated trends detection in social networks.

1. INTRODUCTION

Graphs are used to model many real-world applications such as social networks and biological networks. Nodes in the graph represent objects being modeled and edges represent relationships between nodes. There could be many types of nodes and edges with different attributes in a single graph.

We focus in this work on the problem of detecting geo-correlated trends in social networks [2–4]. Thus, each node in the graph represents either a person in the social graph or a post (*e.g.*, tweet). Geo-correlation leverages the First Law of Geography that states: "Everything is related to everything else, but near things are more related than distant things".

A trend is not a concept that can be defined exactly—trends are manifested in different ways and forms. This leads to many diverse models of trends in social networks [2–4], each with different characteristics. We aim at providing a framework to enable data scientists to explore social network graphs with the purpose of detecting trends. Introducing the human element will provide data scientists with the freedom to infer trends themselves, without the restrictions of pre-defined models. We are encouraged by the success of human-driven data analysis to detect trends [1] and the success of human-assisted methods to search graphs [5].

Social networks are often very large with thousands to millions of nodes and edges. Thus, visual inspection to extract information from the social network graph is an arduous, if not an unattainable, task. Graph summarization techniques are useful to help extracting and understanding information from graph data [6, 7]. Rather than providing statistics to describe graphs, these techniques produce small graphs that preserve the information of the full graph. They also allow the user to control the resolution of these summarization graphs by "drill-down" and "roll-up" operations.

This work leverages the research in graph summarization for the specific application of detecting geo-correlated trends in social networks. We propose summarization methods that are curated to our application.

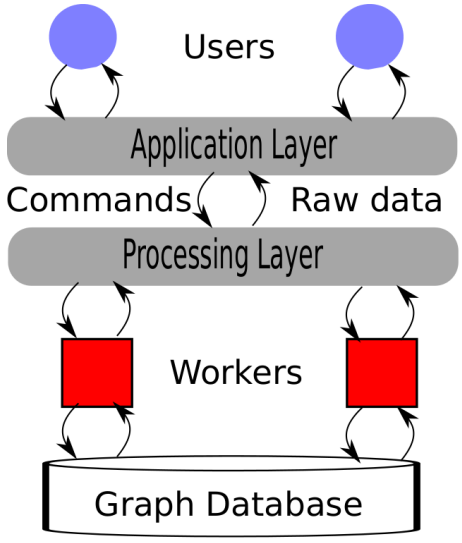

Figure 1: The architecture of the trends detection framework

2. FRAMEWORK

The architecture of our framework consists of four layers depicted in Figure 1. The application layer receives user actions that were performed in the framework's interface. The application layer transforms these actions to meaningful queries and commands to be passed down to lower layers. The application layer also transforms raw data to a presentable format to be displayed back to users. The processing layer is responsible of creating executions to answer users actions. Executions are passed down to workers that query a graph database. Workers cooperate to produce re-

SIGMOD'16 June 26 - July 01, 2016, San Francisco, CA, USA

ACM ISBN 978-1-4503-3531-7/16/06.
DOI: http://dx.doi.org/10.1145/2882903.2914832

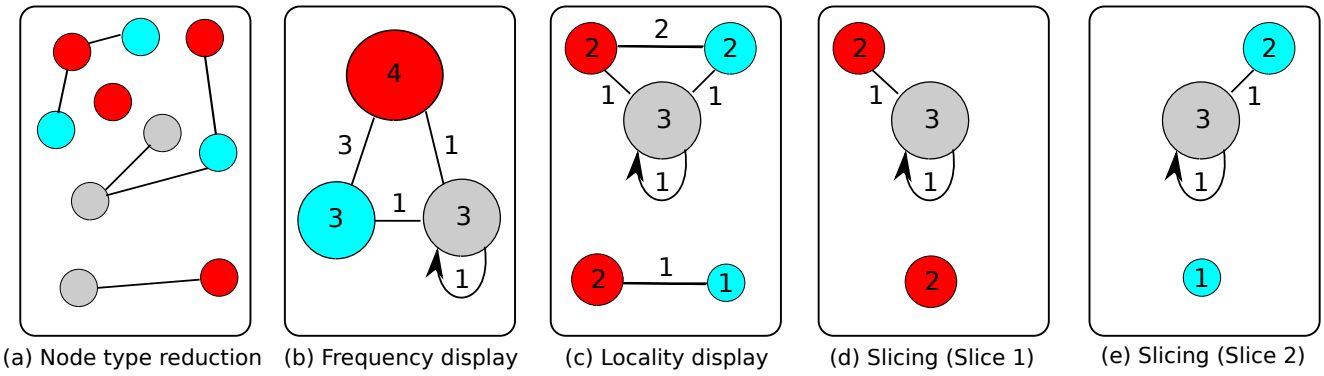

Figure 2: A depiction of the summarization graph after applying each presented method

sults in parallel. And some workers work continuously to produce guiding and predictive information in addition to receiving streams from online social networks to be incorporated in the graph.

The raw graph G consists of nodes and edges. A node represents either a person or a post. An edge exists between two person nodes if there is a connection between them in the social network. Also, a person node is connected to post nodes that the person authored. Nodes and edges have other characteristics and labels that can be manipulated by the user. We focus on the location of posts as they are essential for our geo-correlated trends detection.

3. GRAPH SUMMARIZATION METHODS

Now, we present graph summarization methods that targets the application of geo-correlated trends detection. We demonstrate how the methods affect the summarization graph in Figure 2.

Node type reduction. The full graph contains authors and posts. Our first method, node type reduction, reduces the number of node types. The graph resulting after applying this method contain nodes that represent *topics*. Topics are defined by the users where a post can be mapped to a single or multiple topics. For example, topics can be hash tags mentioned in a post, or they can be the sentiment of the post. In Figure 2(a), for example, the summarization graph has three topics shown by the colors red, blue, and gray, that may represent three hash tags. Nodes capture what the user chose to be the topic of interest. Edges captures the relationship between the authors of posts that resulted in topics. For example, the edge between the red and gray nodes in the bottom means that the posts that resulted in these two topic nodes were authored by persons that are connected in the original graph.

Frequency display. The next method groups topic nodes that are similar to each other. Figure 2(b) shows a grouping of topic nodes that are identical. We call these the frequency nodes. An edge between two frequency nodes represents the edges in the original topics graph. For example, there are three edges between red and blue nodes in the original graph shown as an edge with weight 3 between the red and blue frequency nodes. The user can also control the frequencies that are displayed so that only topics with high frequency are shown.

Locality display. To capture the spatial aspect of topics, we introduce the location display method. This method partitions the graph according to the location of posts. The user has control on what defines a locality by setting a threshold distance—if two posts are within this distance then they are in the same locality. Figure 2(c) shows an example of partitioning the frequency graph into two localities. The top locality represents 7 topic nodes and the lower locality represents 3 topic nodes. Users can then drill down and view different localities individually.

Graph slicing. Graph slicing divides the graph into slices of topics. A slice of the graph contains the nodes representing a subset of the topics. Slices are generated to try to capture topics that might have a relation to each other. This might yield a topic to be in multiple slices. An example of slicing the locality graph in Figure 2(c) is to two slices: slice 1 that contains the red and gray topics (Figure 2(d)) and slice 2 that contains the blue and gray topics (Figure 2(e)).

4. ACKNOWLEDGMENT

This work is supported by NSF grant IIS 1018637, 1528178, and 1442966 and is partially funded by a gift grant from NEC Labs America.

5. REFERENCES

[1] Eighteen hours to thirty six hours entering #tripoli. http://r-shief.org/eighteen-hours-to-thirty-six-hours-entering-tripoli/. 08-29-2011.

[2] C. Budak, D. Agrawal, and A. El Abbadi. Structural trend analysis for online social networks. *Proceedings of the VLDB Endowment*, 4(10):646–656, 2011.

[3] L. Hong et al. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web*, pages 769–778. ACM, 2012.

[4] T. Lappas et al. On the spatiotemporal burstiness of terms. *Proceedings of the VLDB Endowment*, 5(9):836–847, 2012.

[5] A. Parameswaran et al. Human-assisted graph search: it's okay to ask questions. *Proceedings of the VLDB Endowment*, 4(5):267–278, 2011.

[6] Y. Tian et al. Efficient aggregation for graph summarization. In *SIGMOD*, 2008.

[7] N. Zhang, Y. Tian, and J. M. Patel. Discovery-driven graph summarization. In *ICDE*. IEEE, 2010.