SAE: Social Analytic Engine for Large Networks

Yang Yang, Jianfei Wang, Yutao Zhang, Wei Chen, Jing Zhang, Honglei Zhuang,
Zhilin Yang, Bo Ma, Zhanpeng Fang, Sen Wu, Xiaoxiao Li, Debing Liu, and Jie Tang
Deparment of Computer Science and Technology, Tsinghua University
{y-yang-11, jf-wang12}@mails.tsinghua.edu.cn, jietang@tsinghua.edu.cn

ABSTRACT

Online social networks become a bridge to connect our physical daily life and the virtual Web space, which not only provides rich data for mining, but also brings many new challenges. In this paper, we present a novel Social Analytic Engine (SAE) for large online social networks. The key issues we pursue in the analytic engine are concerned with the following problems: 1) at the micro-level, how do people form different types of social ties and how people influence each other? 2) at the meso-level, how do people group into communities? 3) at the macro-level, what are the hottest topics in a social network and how the topics evolve over time?

We propose methods to address the above questions. The methods are general and can be applied to various social networking data. We have deployed and validated the proposed analytic engine over multiple different networks and validated the effectiveness and efficiency of the proposed methods.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Miscellaneous; H.2.8 [**Database Management**]: Data Mining

General Terms

Algorithms, Design

Keywords

Social analytic engine; Social influence; Social network

1. INTRODUCTION

The rapid proliferation of online social networks provides rich data for us to understand the complex mechanism that governs the dynamics of social networks. This has attracted much attention from both academic and industrial communities. For example, SNAP[1] is general purpose network analysis and graph mining library. It is mainly designed for static network and provides several general analytic tools. GraphChi[2] aims to bring web-scale graph computation available on a modern laptop. It supports running very large graph computations on single machine, using parallel techniques. If has focused on computation over graphs, but cannot be directly applied to social networks. For example it ignores several important factors underlying the social networks such as social influence, social status, and structural hole.

In this paper, we propose an Social Analytic Engine (SAE) for analyzing and mining large social network. Figure 1 shows the architecture of SAE. The cornerstone of the analytic engine is a distributed graph database, which provides storage for the networking data. On the top of the database, there are three core components: network integration, social network analysis, and distributed machine learning.

The *network integration* component supports to integrate entities extracted from different networks. For example, in academia, an author may have profile pages on Google Scholar, AMiner, and LinkedIn, but with different accounts. Automatically recognizing and integrating those profile pages can benefit many application such as expert finding and influence analysis.

The *social network analysis* component is our major technical contribution. It first provides basic analyses for network characteristics, including macro-level properties such as density, diameter, degree distribution, community partition, and also micro-level properties such as centrality, homophily, reciprocity, prestige, and reachability for specific nodes. Moreover, we have designed and implemented several novel technologies for social influence analysis [2, 6, 12], structural hole spanner detection [3], and social tie mining [1, 5, 9, 10, 11]. More specifically, social influence aims to quantify the influential strength between users from different angles (topics) in a large social network. Structural hole spanner detection tries to recognize structural hole spanners who control the information flow in the social network. Social tie mining tries to reveal fundamental factors that form the different types of social relationships.

We have also developed a *distributed machine learning* component, which supports to incorporate various social-based factors derived from the above social network analysis component into machine learning models. Employing factor graph model as the example, we demonstrated that

[1]http://snap.stanford.edu/
[2]http://graphlab.org/graphchi/

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

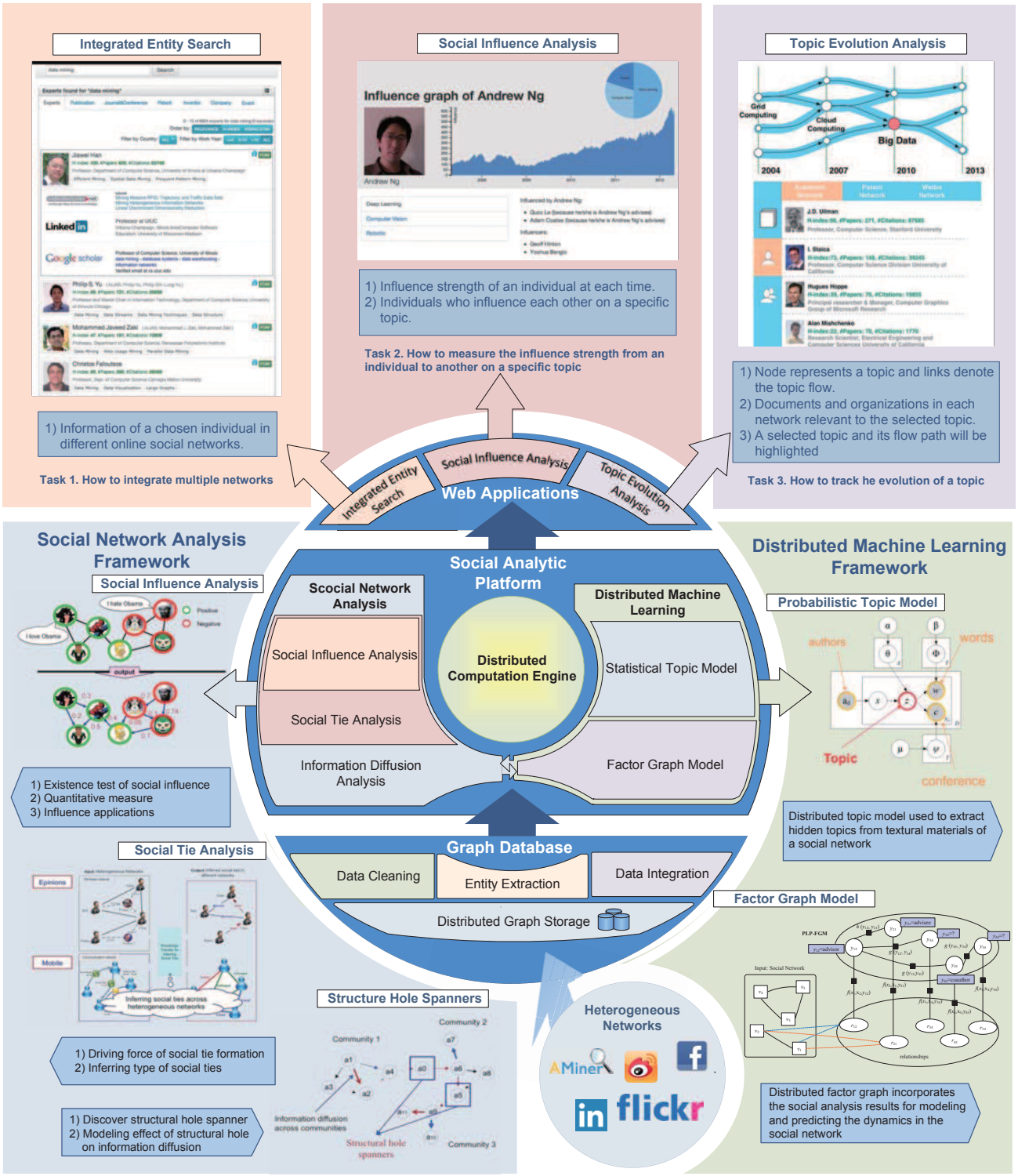

Figure 1: Architecture. We propose an Social Analytic Engine (SAE, http://thukeg.github.io/) for mining large networks. SAE consists of several major components: a graph database, which provides data storage and indexing; a distributed computation engine; a social network analysis component; and a distributed machine learning component.

Table 1: Several typical social networks. #Messages indicates the number of messages/articles associated with the users in the corresponding network.

Data set	#Users	#Relationships	#Messages
Co-Author[3,8,6]	1,629,217	2,623,832	2,174,141
Twitter[1,3]	112,044	468,238	2,409,768
Patent[7,11]	190,000	32,000,000	4,000,000
Weibo[3]	1,787,443	423,347,905	1,038,775,431
Slashdot[5,9]	93,133	964,562	8,714,700
Email[5,9]	151	3,572	136,329
Mobile[5,9]	106	5,436	16,807

by incorporating social theory based features into the factor graph model, the performance of inferring social tie [5] can be significantly improved.

The proposed social analytic engine has many real applications. We will use integrated entity search, social influence analysis, and topic evolution analysis as the example to demonstrate the power of the analytic engine.

2. CORE TECHNOLOGIES

Let us begin with the social networking data we are studying in the analytic engine. Table 1 lists statistics of several typical social networks. Co-Author is an academic network consisting of authors and collaborative relationships between authors. Twitter is a subnetwork crawled using the snowball sampling method. Patent is a heterogeneous network consisting of companies, inventors, and patent full text, derived from a patent database. Weibo is a Twitter like microblogging network.[3] Slashdot is a network of friends and foes. Email is a network derived from the Enron data set, and Mobile is a location based network of mobile users.

We then use three scenarios to introduce the core technologies in SAE: Integrated Entity Search, Social Influence Analysis, and Topic Evolution Analysis.

2.1 Integrated Entity Search

Given a query, finding the most relevant entities (e.g., opinion leaders, authors, and companies) in a network is an important task for social network analysis. The problem has been referred to as entity search or expert finding (when searching for people). However, traditional research ignores one fact: a user may have multiple different accounts in different networks. For example, in the academia, a user may have profile pages in Google Scholar, and AMiner and LinkedIn.

Automatically connecting the entities from different networks can significantly help rank the entities. We first propose a network integration method. The basic idea is that the same entity in different networks would have the same social circles. To implement the method, given two input networks, we first identify potential candidate pairs using heuristics. We then construct a pairwise factor graph by viewing each candidate pair as an observation variable x and associate each pair with a binary variable y to indicate whether the pair represent the same entity. We use a small labeled data set to train the factor graph model by maximizing the log-likelihood objective function $P(Y|X)$ (Y and X are respectively the set of corresponding variables $\{x\}$ and $\{y\}$), and then apply the trained model to identify the other linked pairs.

We have conducted an experiment on Google Scholar, AMiner, LinkedIn, and VideoLecture. With the proposed method, we have accurately integrated more than 10,000 researchers from the four networks.

Based on the integrated results, we propose a propagation-based entity ranking algorithm. Given a query q, the idea is to leverage both textual and network information to calculate a relevance score $r[v]$ for each entity $v \in V$.

More specifically, the method has two stages: in the first stage, we calculate the relevance score of an entity to the query q by utilizing language model; in the second stage, we select the top-ranked entities as candidates and construct a heterogeneous subgraph. We revise the relevance score of candidates by propagating the score between the linked entities in the subgraph. The intuition behind the second stage is, a document written by the user with higher expertise degrees on a topic (query) is more likely to have a higher relevance (or impact), an organization owns higher quality documents and senior/active users should be ranked higher, and a user who writes documents with higher impact should be ranked higher. We perform the score propagation as follows (here we use an organization vertex c as the example):

$$r^{k+1}[c] = (1-\xi_1-\xi_2)r^k[c] + \frac{\xi_1}{|V_a^c|}\sum_{d\in V_a^c} r^k[a] + \frac{\xi_2}{|V_d^c|}\sum_{d\in V_d^c} r^k[d]$$

where $r^k[c]$ is the ranking score of organization c after the k-step propagation; V_a^c denotes a set of users related to organization c and V_d^c denotes a set of documents owned by organization c; ξ_1 and ξ_2 are two parameters to control the propagation. The number of propagation steps reflects how we trust the network information. Setting $k = 0$ indicates that we only use the content information, while setting $k = \infty$ indicates that we only trust the network information, thus the algorithm obtains a result similar to that of PageRank [4] on the heterogeneous network.

2.2 Social Influence Analysis

Social influence occurs when one's opinions, emotions, or behaviors are affected by others. It is widely viewed as a fundamental force that governs the dynamics of social networks. In SAE, we provide a unique feature to estimate users' topic-based influence. As the example given in Figure 1, Quoc Le is identified as one of the major influencees of Prof. Andrew Ng.

The input of this problem is a social network $G = (V, E)$, and a T-dimensional topic distribution $\theta_v \in \mathbb{R}^T$ associated with each vertex (user) in G. Each element θ_v^z is the probability (importance) of the user on topic z, satisfying $\sum_z \theta_v^z = 1$. The topic distribution is obtained by topic modeling methodologies in the distributed machine learning component of SAE. The goal of social influence analysis is to derive the topic-level social influence based on the input network and topic distribution on each user. Formally, social influence from user v to v' denoted as $\mu_{vv'}^z$ is a numerical weight associated with the edge $e_{vv'}$ and topic z. In most cases, the social influence score is asymmetric, i.e., $\mu_{vv'}^z \neq \mu_{v'v}^z$, and the social influence from user v to v' will vary on different topics. We also define one's global influence strength as $\mu_v = \sum_{v',z} \mu_{v,v'}^z$.

[3] http://weibo.com

To quantify the influence between users, it is necessary to consider both user-specific topic distribution and the network structure. We employ a unified approach to utilize both the local attributes (topic distribution) and the global structure (network information) for social influence analysis. The algorithm is called as TAP (topical affinity propagation) [6]. The main idea is to perform affinity propagation at topic-level for social influence identification. The approach is based on a factor graph model, in which the observation data are cohesive on both local attributes and relationships. To calculate the influence in different periods, we execute the proposed method on networks at each time point respectively.

2.3 Topic Evolution Analysis

Understanding what are the hottest topics and how the topics evolve over time is an important and challenging task. SAE addresses this problem by supporting to construct a *topic flow graph*. In a topic flow graph, a vertex denotes a topic, and an edge represents a topic flow. A topic flow from one topic to another at adjacent time points represents the former one evolves (splits or merge) into the latter one. As the example shown in Figure 1, the "big data" topic is very hot recently and the topic can be traced back to "database" and "parallel computing". A vertex without any in-edges represents an emerged topic. A vertex without any out-edges denotes a topic that people do not talk about. The strength of the arrow between two vertices indicates the similarity of the two corresponding topics. The size of a vertex represents the hotness of the corresponding topic.

Given a query, SAE returns a topic flow graph containing relevant topics to the query. The user is allowed to choose a vertex (topic). The system then identifies the backbone path containing the chosen topic, and presents all topics in that path. The system also displays a table to present documents, users, and organizations relevant to the selected topic in each social network.

The topic flow graph can be defined as a directed acyclic graph (DAG), where vertices represent the topics at each time point, and edges indicate the topic flow between two consecutive topics along time. Given a set of documents over time, technically, the goal is to extract hidden topics and construct a DAG which represents the flow of topics.

To automatically generate the topic flow graph for a given query, we first utilize the probabilistic topic model implemented in the distributed machine learning component of SAE to extract hidden topics from the documents at each time point independently. Each topic is assigned with a word distribution ϕ_{zw}, which indicates the probability that the word w will be used to express the topic z. We then employ classic clustering algorithms like K-means to cluster topics in consecutive time points. We define the distance between two topics as the KL-divergence of their word distribution. For two consecutive topics along time, we create an edge if their distance is less than a threshold, and define the strength of the edge to be inversely proportional to the distance.

3. CONCLUSIONS

In this paper, we present the Social Analytic Engine (SAE) for large networks. We briefly introduce the architecture of the system and then use three scenarios as the example to explain the core technologies: integrated entity search, social influence analysis, and topic evolution analysis. The engine also provides many other social network analysis tools such as kernel community detection, structural hole spanner detection, and user behavior modeling. We deploy the SAE engine over a number of different genres of data sets and our studies validate the effectiveness and efficiency of the proposed social analytic engine.

Acknowledgements. The work is supported by the Natural Science Foundation of China (No. 61222212, No. 61073073, No. 61170061), Chinese National Key Foundation Research (No. 60933013, No.61035004), and a research fund supported by Huawei Inc.

4. REFERENCES

[1] J. Hopcroft, T. Lou, and J. Tang. Who will follow you back? reciprocal relationship prediction. In *CIKM'11*, pages 1137–1146, 2011.

[2] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang. Mining topic-level influence in heterogeneous networks. In *CIKM'10*, pages 199–208, 2010.

[3] T. Lou and J. Tang. Mining structural hole spanners through information diffusion in social networks. In *WWW'13*, pages 837–848, 2013.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford University, 1999.

[5] J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogeneous networks. In *WSDM'12*, pages 743–752, 2012.

[6] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *KDD'09*, pages 807–816, 2009.

[7] J. Tang, B. Wang, Y. Yang, P. Hu, Y. Zhao, X. Yan, B. Gao, M. Huang, P. Xu, W. Li, and A. K. Usadi. Patentminer: Topic-driven patent analysis and mining. In *KDD'2012*, pages 1366–1375, 2012.

[8] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *KDD'08*, pages 990–998, 2008.

[9] W. Tang, H. Zhuang, and J. Tang. Learning to infer social ties in large networks. In *ECML/PKDD'11*, pages 381–397, 2011.

[10] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo. Mining advisor-advisee relationships from research publication networks. In *KDD'10*, pages 203–212, 2010.

[11] Y. Yang, J. Tang, J. Keomany, Y. Zhao, Y. Ding, J. Li, and L. Wang. Mining competitive relationships by learning across heterogeneous networks. In *CIKM'12*, pages 1432–1441, 2012.

[12] J. Zhang, B. Liu, J. Tang, T. Chen, and J. Li. Social influence locality for modeling retweeting behaviors. In *IJCAI'13*, 2013.