Reaction Paper
Data-driven Outbreak Detection in Social Networks

Jiayuan Ma
Xincheng Zhang
Pearl Tan

October 18, 2012

1 Introduction

Recently, the process of how ideas and influence propagate through a social network has drawn extensive research efforts. It lies in the core of understanding the diffusion of news, fashions and innovations, it also plays important roles in marketing strategies, virus and pollution control. This reaction paper focuses on: (i) summarization of literatures in influence maximization and outbreak detection, (ii) discussion of pros and cons of existing models and algorithms, and (iii) proposal of what we will research in our projects.

2 Literature Survey

Since the study of influence spread in networks has a long history, we are not able to cover every aspect of the field in this reaction paper. Here, we will mainly summarize papers on information maximization and outbreak detection.

2.1 Two Important Problems

Before getting started, we define the following two important problems that we will study in this reaction paper.

> **Influence Maximization** (IM): Given a social network, with weights on the edge representing the extent to which individuals influence one another. We'd like to trigger a cascade of information diffusion by selecting a few (usually a given number) seed nodes in the network. The problem of influence maximization asks how we should choose the seed nodes so as to maximize the spread of influence (in terms of the number of influenced people).
>
> **Outbreak Detection** (OD): Suppose that there is a dynamic process spreading over a social network, for instance virus spreading over a contact network, polluted water over a distribution network, news over a micro-blog network, etc. Outbreak detection asks us to select a set of nodes to detect the process as effectively as possible (in terms of the detection time, or the number of effected people). Under Leskovec et al. [8]'s proposed model, IM is a special case of the network outbreak detection problem.

2.2 Basic Diffusion Models

In Kempe et al. [6]'s highly-cited paper, two basic information diffusion models have been generalized, which are Independent Cascades (IC) model and Linear Threshold (LT) model. Both of them are the most widely-studied diffusion models today.

Given a graph $G = (V, E, w)$, where V is a set of n nodes, $E \subseteq V \times V$ is a set of m directed edges, and $w : V \times V \longrightarrow [0, 1]$ is a weight function such that $w_{u,v} = 0$ if and only if $(u, v) \notin E$. We start the diffusion process with an initial set of active nodes S_0, and the process cascades in discrete steps $i = 0, 1, 2, \ldots$. Let S_i denote the set of vertices activated at step i. The whole process stops at t when $S_t = \phi$. IC and LT only differ in how every individual node is activated, and we will explain their differences.

Independent Cascades (IC): If a node u first become active in step i, it is given a *single* chance to activate each of its inactive neighbor v, with the probability of success being $w_{u,v}$. Each successfully-activated node v will become active in step $i + 1$. Notice that this process is *unrepeatable*: we cannot make any further attempts on the same edge.

Linear Threshold (LT): An LT influence graph further assumes that $\sum_{v \in V} w_{u,v} \leq 1$ for every u. The dynamics of LT proceed as follows

Each node u has a threshold θ_u which is uniformly distributed in the interval $[0, 1]$, which models the uncertainty of individual's conversion threshold.

The node u is activated when the weighted sum of its activated neighbors v is no less than the threshold θ_u. Mathematically, node u will become active if

$$\sum_{v \in \cup_{0 \leq j \leq i-1} S_j} w_{u,v} \geq \theta_u \tag{1}$$

Both the two models have, to some extent, simulated the way how information spreads. The IC model treats the spread of information independently and probabilistically, where information spreads through edges independently and with a given probability. The LT model focuses more on collaborative and threshold side of influence propagation, where sufficient number of neighboring influence will help spread the information. Notably, both of these two models are progressive, which means that nodes can switch from being inactive to being active, but do not switch in the other direction.

2.3 Related Work

Table 1 summarizes some related past papers w.r.t what model the paper suggests, what problem the paper is trying to solve and what algorithm the paper proposes in its context. Here, we summarize the main idea and contribution of each paper

Kempe et al. [6] generalized the IC and the LT model from the previous research, and formulated influence maximization into a NP-hard optimization problem. The authors utilized an analysis framework based on submodular functions, and proposed an efficient greedy algorithm that approximates the optimal solution with a provable bound.

Paper	IC	LT	IM	OD	Algorithm
Kempe et al. [6]	√	√	√	×	greedy algorithm
Leskovec et al. [8]	√	×	×	√	greedy algorithm with lazy forward optimization
Chen et al. [2]	√	√	√	×	1. greedy algorithm on pre-generated graphs 2. degree-discount heuristics
Chen et al. [3]	×	√	√	×	1. linear time algorithm on DAGs 2. LDAG for non-DAGs

Table 1: Summary of related work

Leskovec et al. [8] studied the problem of outbreak detection and demonstrated that many outbreak detection objectives are submodular functions. The authors combined the original and the normalized greedy algorithm to get an algorithm which effectively approximates the optimal solution. Furthermore, the authors proposed the CELF algorithm, which exploits submodularity to speed up the computation by 700 times.

Chen et al. [2] focused on improving the time efficiency of [6] and [8] in very large networks. Their contributions are two-fold: (i) improvement of greedy algorithms in IC models by pre-generating the graph (ii) proposal of novel degree discount heuristics that is cheap to compute (in milliseconds) and outperforms traditional degree and centrality heuristics.

Chen et al. [3] proposed the first influence maximization algorithm specifically designed for the LT model. The authors proved #P-hardness of IM under the LT model in general graphs, proposed a linear-time algorithm for DAGs, and applied the proposed linear-time algorithm to local DAG structures in general graphs.

Goyal et al. [5] proposed an solution framework for regression the previous propagation action and time factor into the LT model weight parameter. The evaluation result shows time variant model performs a lot better than static one.

3 Further Discussions

3.1 Submodularity

We observed that mainstream of the past research papers [6, 8, 2, 3] formulate their problems into a general submodular framework, where the objectives to optimize are submodular functions. Under this framework, the following two results can be attained.

Maximization of the objective functions is a NP-hard problem.

Greedy hill-climbing algorithms give an approximation of at least $(1 - \frac{1}{e})$ of the optimal.

While the tight bound of greedy algorithms gives submodular formulation a great advantage over other formulations, the intrinsic hardness of this formulation still makes the computation prohibitive. Although Leskovec et al. [8] exploited "diminishing returns" of the submodularity to

speed up the computation, [6, 8, 2, 3] only uses submodularity to justify the effectiveness of the greedy algorithms. It is quite difficult to utilize the property of submodularity. Submodularity may be a too weak property for this problem, we can suggest a stronger property of the function (e.g. convex and concave) to improve over both effectiveness and efficiency.

3.2 Blind Point in Existing Models

Independent Cascade and Linear Threshold Model are built up for simulating the influence spread process in the network. However, they both have disadvantages in capturing the real social network propagation. Independent Cascade presumes that neightbors activated the node independently. In social network situation, however, the activation process is not actually independent. Especially, in a small world situation like social network, the high clustering coefficient may lead neighbours referring each other's action to actually fire an adoption. Linear Threshold based on the graph statistic may neglect the fact user will adopt an action by its content. For example, my neighbours can be all interested in one topic while I am not. It leads to the fact that my activation chance won't get increase when multiple neightbours get activated.

3.3 Degree and Centrality

Degree and centrality play crucial roles in both outbreak detection and influence maximization problem. Intuitively, nodes with large degree have more chances to spread the information around, and nodes with small distances to the others are more likely to spread the information quickly. In the scenario of outbreak detection, it seems that nodes with larger degree will certainly help the spread process. Nonetheless, if the large degree node tends to get activated/polluted among last several nodes in cascade, it is not that meaningful to put sensor on the large degree node. However, the result of Kempe et al. [6] revealed that their proposed greedy algorithm outperforms traditional degree and centrality heuristics. Chen et al. [2] proposed degree discount heuristics which are super fast to compute and has comparable performance to the greedy algorithms. We believe that degree is still an highly important factor in the outbreak detection problem and we will have to trade off between degree-based heuristics and greedy algorithms to get both efficiency and effectiveness.

3.4 User Influence

Unlike water pollution problem, every spot gains approximately equal chance to affect their neighborhood, an user in a social network is influenced differently by different users. Cha et al. [1] discovered that influential power tends to be higher locally instead of globally in Twitter. However, we believe that the influence of the user has not being well studied enough. In Goyal et al. [5], they use machine learning technique to correlate the previous propagation into the influential model parameters. It is obvious that previous adoption experience is a strong indicator of high influential power. Not satisfying with this result, we tried to analyze and get what factors contribute to influential power instead of only examine the outcome of high influential power.

3.5 Content Utilization

Through reading the existing approaches for inference maximization and outbreak detection, an interesting phenomenon has been observed that all of approach only consider exploiting the graph property of the user social/relation network. The property being considered in these papers are

limited to the degree, position and some basic property of a node in a graph. However, we believe that the thing will become convoluted in the social network graph due to the flow is information, neither a new type of virus of polluted water. Intuitively, user in a social network have relatively strong subjective idea to choose if they will adopt an idea or not. Nonetheless, it seems that the content is not being taken into consideration in the influence maximization problem. Moreover, the outbreak detection tried to placed the sensor in the intermediated node which may tends to have high degree. In social network scenario like tweeter, it may be already late to discover the rumor spreading process since people may passive adopt the information.

3.6 Heterogenous Information Diffusion

In Leskovec et al. [7] paper, an easy and elegant model has been introduced to show the temporal dynamics of various topic flows in blog space. It is worth to notice that unlike outbreak of a virus or water pollute situation, various topic can flow in social network. Therefore, we believe there is a hidden categorization behind the information flow in social network. If we represent each edge with weight representing influence, the weight will change accordingly with the topic currently propagate. Therefore, instead of a static network model to simulate a propagation, we will try to come up with time and content dependent network model.

4 Brainstorming

We believe there are already observable amount of work has been done in improving the submodular function computation both in Independence Cascade and Linear Threshold Model. Therefore, our potential research direction will be try to introduce the content into the influence maximization problem in following several scenario.

4.1 Graph Construction

Previously, the graph construction and the linear threshold model weight setup are only taken consideration of the node information. We should try to build regression model to correlate the keyword, meme, or phrases into the weight system. Goyal et al. [5] enables the approach of using machine learning technique to learn the linear threhold probability of activation. Instead of only considering the factor from node itself and propagation, content will be contributed to the activation process. A statistic of the phrases or word will be setup in order to modeling the user interest on topic. Apart from topics in the information content, user will be also considered as a topic. For instance, twitter user tends to re-tweet whatever content their star post on the timeline. Machine Learning technique will be employed to do the modeling and construct the graph for a given piece of information. Therefore, depends on different type of information, the graph varies and a better and earlier outbreak may be detected even the piece of information has not been published by those node with high degree.

4.2 User Interested Topic Inference

Instead of discovering the outbreak, this scheme can also detect the usertopic relation of which categories of phrases, keywords or even users are contributing to the information outbreak more than others. Since statistic regression will be constructed through training the topics to graph

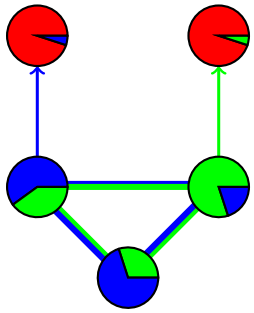

Figure 1: Topic Inference: the three users in the triad share the same interest 'blue' and 'green', therefore the edge between them are in both direction and both blue and green exist, however, the two 'outsider' only have small portion of interesting in blue and green. They still get small chance to be activated by nodes in the triad.

construction, those results can actually be used to infer which topics are certain users interested in. Moreover, we can potentially get categories easier to trigger the information cascade than others.

5 Project Proposal

Based on our previous literature reviews and discussions, we will focus our project on injecting information content into the diffusion network, which will hopefully help information maximization and outbreak detection.

5.1 Problem Formulation

Generally, we will perform outbreak detection in content-rich social networks, where there is a sufficient amount of content attached to each node in the graph. We will utilize a data-driven approach to construct the influence graph, automatically learn the edge weights, and run outbreak detection on the generated graph.

5.2 Injecting Content into Models

We will inject information content to both IC and LT models.

Adapted Independent Cascade (IC) Model In IC model, each neighbor v of node u will have a single chance to activate u with probability $w_{v,u}$. Instead of setting the probabilities solely based on the degree statistics, we will automatically learn these probabilities from the graph structure and the information content. We will use features such as in-degree/out-degree of the node, the content information previously spread by the node, and node's subjective preference, etc. By employing discriminative learning algorithms, we will use a data-driven approach to figure out what is the best probability distribution for activation.

Adapted Linear Threshold Model (LT) Model In LT model, we will use the regression model to figure out for different content and different user, how much weight should be given

to w and whether we should adjust the threshold for θ. For instance, users may have a lower activation threshold for the topics they are interested in, and users with similar interests will be structurally grouped together.

5.3 Topic Inference

Since data-driven approach and discriminative algorithm will be employed in the development of our probabilistic graph construction, the probabilistic parameter in our model will indicate the correlation between topics and users. As visualized by Fig. 1, we tried to capture the topics one user are interested in as well as flows of topics between users. The visualization result is actually overlaying graphs generated under different topics by our model.

5.4 Data Set

Twitter is a social network where nowadays even journalist are searching news from. The information outbreak in Twitter happens frequently. Therefore, detecting which type of information tweeted by which type of user will have a higher chance to spread in the network is a potentially valuable problem. The result can be both used for provide right advertisement propagation or detect the rumor earlier. However, as it has been examined in the previous research [4], twitter may have relatively small cascade and the noise in twitter is very large. A data refinement scheme has to come up to minimize the noise and exploited the cascade as much as possible. After brainstorming, we think the collaboration network maybe a much cleaner dataset to start with as a proof of concept step. Each paper will have keyword so there is no much need to clean the noisy content from the collaboration network. And investigating how a new topic is spreading out through the academic social network will also be meaningful.

5.5 Technical Challenge

The refinement of Twitter content may meet some unpredictable challenges as filtering out the spam information and dealing with not English word. Also, we have to setup a ground truth in Twitter or the collaboration network to judge whether a piece of information or a new research topic has been outbreak through the network.

The major technical challenge will be choosing an appropriate machine learning model to project the feature from content onto the outbreak prediction graph. Instead of using complicated model, we will first try out several simple but useful model such as Naive Bayes as a starting point to see how the phrases and influenced people contribute to the influential process.

5.6 Evaluation

After the model have been built up, we can utilize it to predict a propagation cascade by properly selecting originating node. The ground truth we pick here will be the real propagation cascade. We will try to compare the activation nodes overlap and cascade size to evaluate the model. In the meantime, the model will be applied to choose where to place outbreak detection sensor, and compare the result to the existing algorithm. Apart from predicting propagation cascades, we will also infer different topics user interested in and visualize them in the graph like Fig. 1.

References

[1] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *4th international aaai conference on weblogs and social media (icwsm)*, volume 14, page 8, 2010.

[2] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208. ACM, 2009.

[3] W. Chen, Y. Yuan, and L. Zhang. Scalable influence maximization in social networks under the linear threshold model. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 88–97. IEEE, 2010.

[4] S. Goel, D. Watts, and D. Goldstein. The structure of online diffusion networks. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 623–638. ACM, 2012.

[5] A. Goyal, F. Bonchi, and L. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM, 2010.

[6] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[7] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM, 2009.

[8] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429. ACM, 2007.