Influence Analysis in Business Social Media

Flora Amato[1], Vincenzo Moscato[12], Antonio Picariello[12], Giovanni Ponti[3] and Giancarlo Sperlì[1]

[1] Dip. di Ingegneria Elettrica e Tecnologie dell'Informazione, University of Naples "Federico II", Naples, Italy
[2] CINI - ITEM National Lab Complesso Universitario Monte Santangelo, 80125, Naples, Italy
[3] DTE-ICT-HPC - ENEA - C.R. Portici P.le E. Fermi, 1 80055 - Portici (NA), Italy
{flora.amato, vmoscato, picus, giancarlo.sperli}@unina.it, giovanni.ponti@enea.it

Abstract. In this paper, we describe a novel data model for particular online business social networks such as Tripadvisor and Yelp: we also define a greedy influence maximization algorithm to determine the most influential users on the base of proper influence patterns. The result of such analysis is then combined with some economic data in order to propose a set of possible financial strategies for business objects. Finally, a case study and some preliminary and interesting results are presented for the Yelp dataset.

Keywords: Influence Analysis, Social Network Analysis, Influence maximization

1 Introduction

Nowadays, there is a massive usage of social networks, and this phenomenon affects many contexts and real life scenarios. People express opinions and sentiments, often sharing also multimedia contents. The diffusion of these new communication instruments allows the proliferation of a big amount of data, and Internet becomes the most important source of social network data.

However, handling and analyzing such data is not a trivial task, since several aspects should be taken into account. First of all, social network databases are very complex and heterogeneous, storing different kinds of information. Secondly, the huge amount of available data and stored information makes it hard to analyze and correlate, falling into the field of *big data*. Another important challenge consists in the problem of applying traditional data mining techniques on social data in order to extract information and correlations.

In this paper, we focus on the field of social networks in which users express opinions, judgments, and ratings about items, places, and/or services. There are lots of platforms allowing to do this, and Tripadvisor and Yelp are only the most known and representative ones.

In particular, we aim to investigate the problem of how one user can influence other user judgments with her/his opinions and reviews on an item. Since the

items objective of the judgments belong to the context for which the social network has been thought, it can also contains economic data. In this case, we refer to a particular type of items, called *business objects*.

Analyzing business objects in a social network context should take into account not only object details and economic aspects, but also the influence and the bias that user judgments give to other users. In this direction, the goal of our work is twofold: in a first phase, we propose a model for the social network allowing to discover the most influential users and, secondly, how this result can be merged with economic data in order to propose financial strategies for business objects.

In order to understand the potentiality of Social Network Analysis for financial purposes, let us suppose that we are in a big city area in which there are several types of business objects (such as pubs, museums, hotels and so on) that provide different services and/or products. We can associate to each business object several information regarding their main features (i.e. open hours, atmosphere, types of cousin, etc.), and also economic information (i.e. ticket information, menu price, etc.). Users are able to make several reviews, that are short text descriptions of their experience in these firms after they visit them. An automatic system could perform a NLP analysis of the reviews, in order to extract the sentiment of each one, and successively derive an homogeneous graph, obtained by correlating temporal and semantic information, to infer weighted relationships between users. Exploiting the features and the properties of these graphs, we identify the most influential users focusing on a specific area or on a given category of business objects. The retrieved social influence may be used in viral marketing: the small subset of influential users in the social networks is essential for marketing a product, exploiting the attitude of the most influential people for suggesting to a given business object a specific advertising strategy, a different type of approach to hit these specific users or to create alliance with other shops in order to increase their revenue growing up the amount of users attending their business activity.

The rest of the paper is organized as follows. Section 2 discusses the state-of-the-art for the influence maximization and other research issues in business social networks. In Section 3, the proposed model for social network and the building of the influence graph are described, whereas Section 4 discusses the adopted influence maximization strategy. Section 5 shows preliminary experimental results. Eventually, Section 6 gives some conclusions and discussions of the paper.

2 Related Work

The technological development leads to see Online Social Networks (OSNs) an important means on which it is possible to make business, allowing firms to promote their products and services and users to provide feedbacks about their experiences.

In [13], the authors propose an approach for rating prediction based on matrix factorization that exploits both the rich attributes of items and social links of users to address the cold-start problems. In particular, the *Kernel-based Attribute-aware Matrix Factorization* (KAMF) is developed to discover the non-linear interactions among attributes, users and items. Further, it is provided an extension of KAMF approach to address the cold-start problem for new users by utilizing the social links among users.

Nowadays, the continuous increase in the use of smartphones has led to the emergence of a new type of social network called Location Based Social Networks (LBSNs); the peculiarity of these networks is to integrate the user's location into the well-known OSN capabilities, introducing new research topics and issues.

An influence maximization problem [14] proposes to maximize the benefit of location promotion, leveraging the feature of LBSN.s To address this problem, the authors propose two models to capture the check-in behavior of individual LBSN users, based on how location-aware propagation probabilities can be derived.

Indeed, a large number of people rely on content published on OSN for making their decisions. This facet leads to engage spam reviewer to increase the popularity of particular business objects or to harm a possible competitor.

Thus, the spammers detection is another important research topic in a LBSN; in fact the 20% of the reviews in the Yelp website are actually spam, as shown in [1]. Shehnepoor et al. [9] propose a novel framework, named *NetSpam*, which utilizes spam features for modeling review datasets as heterogeneous information networks to map spam detection procedure into a classification problem in such networks.

The wellness of information in OSNs has allowed to focus on *Viral Marketing* applications, developing new marketing strategies based on "word-of-mouth". Different models have been proposed in literature to describe the spread of idea in OSNs.

A family of approaches are based on *Stochastic models*, in which a node can be in an *active* or *inactive* state depending on the influence exerting by its activated neighbors. Granovetter and Schelling [5] propose the first models based on the threshold for each node to represent the spread of idea or innovation in a social networks. In this family, the main models are *Linear threshold* (LTI) and *Indipendent Cascade* model. In the first one a node is activated if the weighted sum of the activated neighbor is greater than a node-specific threshold, while in the second one each activated node have one chance to activate its neighbors. The spread of influence can be also modeled by epidemiological models, in which the process is described as an infection that evolves in the biological population. An example of these category is the mode propsed by l [8], in which a user could be in two state susceptible or infected.

Based on the spread models, the influence maximization problem has been proposed in literature. A first work on this topic is provided by Richardson and Domingos [3] deal with a first problem of influence maximization related to Viral Marketing application. In particular, they examine a particular market as

a social network composed by different interconnecting entities and model it as a Random Markov field in order to identify a subset of users to convince to adopt a new technology for maximizing the use of it. The choice of the most influence nodes is an optimization problem that has been proven by Kempe et al. [7] to be NP-Hard. A first approach [7] is based on Montecarlo simulation that exploits hill-climbing strategy. This approach leverages sub-modular influence function that allows to obtain a solution that is no worse than $(1-1/e)$ of optimal solution. Moreover, the Montecarlo simulation provides a more accurate approximation of influence spread, which turns out to be inefficient for large networks. To overcome the inefficiency of Montecarlo simulation two different approaches have been proposed: *heuristic based* methods [6, 12] exploit communities or linear systems for restricting the influence spread, while *sketch based* techniques [2, 10] build a family of vertices sets exploiting the reverse simulation.

3 Model

To model a real market, we define a particular network, called *Knowledge Sharing Network*, in which there are $N \geq 2$ firms that provide several services and M users. In particular:

Definition 1 (Knowledge Sharing Network) *A* Knowledge Sharing Network *(KSN) is a quadruple $G = (V; E; \gamma; \omega)$, where V is a set of vertices, E is a set of edges, and γ and ω are two weight functions such that $\gamma : V \rightarrow [0, 1]$ and $\omega : V \times V \rightarrow [0, 1]$ The set of vertices is defined as $V = U \cup BO$, with* U *being a set of* KSN *users and* BO *a set of business objects. The set of edges is in turn composed by three kinds of relationships:* Social relationships*, established between two users,* Business relationships*, existing between a user and business object, and* Neighbor relationships*, established between two business objects that are located in the same geographic area and belong to the same category.*

In our vision of KSN, users and business object nodes are generic nodes containing several attributes such as preferences, price, business info, open hours and so on. Thus, a *KSN* model allows to represent heterogeneous entities and relationships in a unified network.

Example 1 (Example of KSN) *To better explain our idea we show an example in Figure 1. Let $Vinni$, $Giank$, $Flora$ and $Picus$ be four users and BO_1, BO_2 and BO_3 be three firms. The four users made reviews in different times ($t_1 > t_2 > t_3 > t_4$) about the three business objects, two of which are linked by a neighbor relationship. It is possible to note how it is easy to represent different types of relationships by our model. Eventually, each user or business object is respectively composed by several attributes about preferences or services.*

The influence analysis problem seeks to identify the best subset of users that allows to maximize the spread of new products, new idea or so on. Moreover, in the real world, there are different firms competing among themselves with the

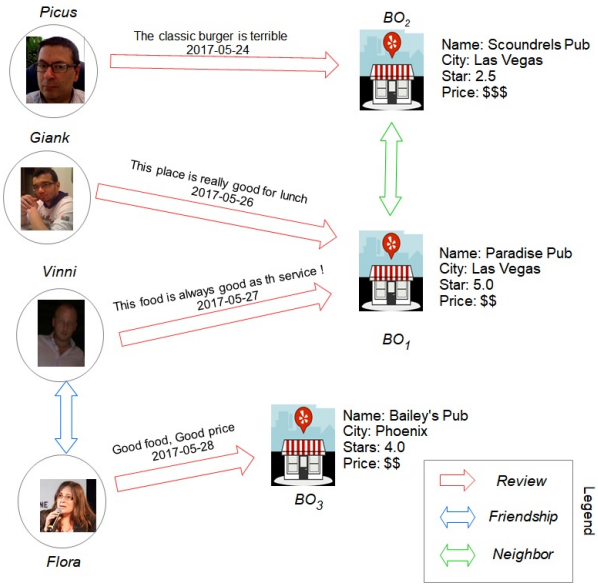

Fig. 1: An example of KSN

aim to maximize its profit in order to attract an increasing number of customers. In particular, the influence exerts from a user respect to another one is computed by evaluating several common actions made on the same business object.

To analyze user behavior we exploit the concepts of *social path* and *relevant social paths*.

Definition 2 (Social Path) *A* social path *between two vertices* v_{s_1} *and* v_{s_k} *of a KSN is a sequence of distinct vertices and edges* $sp(v_{s_1}, v_{s_k}) = v_{s_1}, e_{s_1}, v_{s_2}, ..., e_{s_{k-1}}, v_{s_k}$ *such that* $\exists e(v_{s_i}, v_{s_{i+1}}) \in E$ *for* $1 \leq i \leq k-1$. *The* length λ *of the path is* $\alpha \cdot \sum_{i=1}^{k-1} \frac{1}{\omega(e_{s_i})}$, α *being a normalizing factor. We say that a social path* contains *a vertex* v_h *if* $\exists e(v_i, v_h) : e \in sp(v_{s_1}, v_{s_k})$.

The above definition allows us to correlate the path length with the concept of distance between two nodes interconnected by the examined path; exploiting this definition, we can perform pruning strategies based on a given threshold. The following example will better explain our basic ideas.

Example 2 (Social Path) *Considering the network in Example 1, we can identify several social paths, representing communication flows between users by social channels. In particular, the business objects is a communication means to interconnect users in this type of network. In Figure 2 we represent in concise way these types of information.*

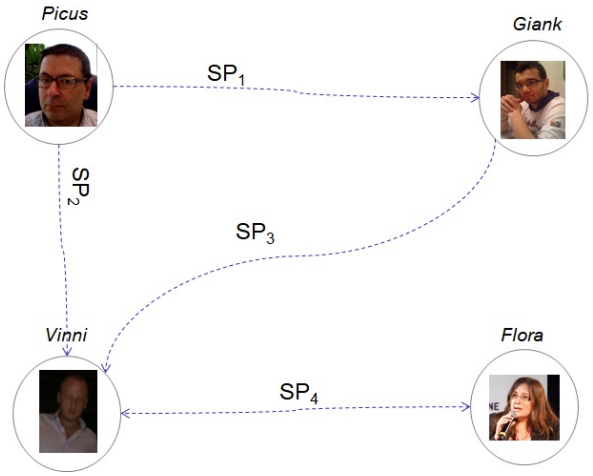

Fig. 2: An example of social paths

Definition 3 (Relevant Social Path) *Let Θ be a set of conditions defined over the attributes of vertices and edges of a KSN. A* relevance social path *is a social path satisfying Θ.*

A particular kind of relevant social path is constituted by the *influential paths* that connect two users and by which a user can "influence" other users. It is possible to identify *weak* and *strong* influential paths; in particular, in the first case, the user u_i influences u_j, if the user u_j posts a review of the same sentiment of the review previously posted by u_i on the same business object, while in the second one two users have to be also friends.

Example 3 (Influential Path) *Considering the Example 2, we can identify three relevant social paths. Each social path corresponds to a sequence of reviews or tips following given rules; for instance, a social path is instantiated due to the fact that* Giank *made a review on BO_2 before to* Vinny *and with the same sentiment, while another path is instantiated due to the fact that* Picus *made a review on another business object close to BO_2 with an opposite sentiment.*

To deal with the influence maximization problem, we build an homogeneous graph (*Influence Graph*) $IG = (V; E; \omega)$ whose vertices are specific users of KSN; in particular, there exists an edge e between two vertices v_i and v_j for all influential paths connecting v_i and v_j. For each edge, the related weight will be determined as in the following:

$$\omega(e_{i,j}) = \frac{\sum_{k=1}^{M} \gamma(sp_k(v_i, v_j))}{N_j}, \tag{1}$$

M being the number of distinct influential paths between v_i and v_j and N_j the number of influential paths of having as destination vertex v_j.

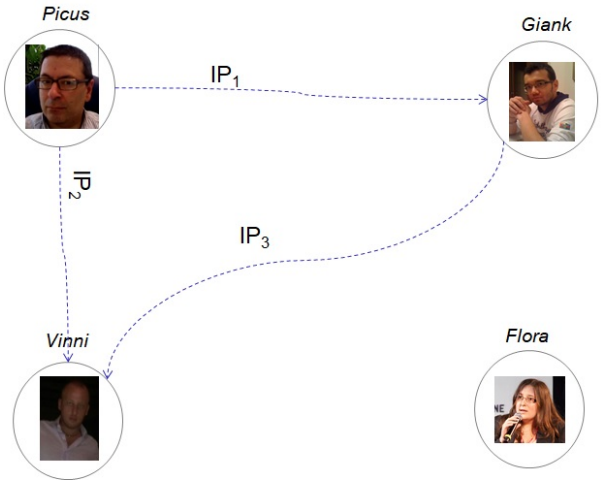

Fig. 3: Example of relevant paths

Example 4 (Influential Graph) *Considering the Example 3, we can build the Influence Graph, shown in Figure 4, by exploiting the equation 1. In particular, it is possible to note that* Picus *can influence two users:* Giank *and* Vinny*, since they could have seen the previous negative review made by* Picus *on* BO_1*, that is a business object close to* BO_2*.*

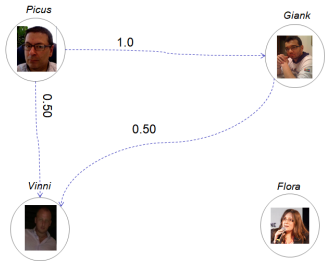

Fig. 4: An example of Influence graph

4 Influence Maximization Problem

The influence maximization problem aims at identifying a minimum number of users that maximize the spread of new idea or technologies in the OSN. Nevertheless, given that the interest of users for a specific product or business object is variable due to different causes, for instance the change of opinion of friends about the examined products or the introduction of new product, the

temporal dimension acquires an increasing value. The scope of firms' market campaign is to maximize the spread of new product in a specific time interval, for instance couple of weeks or few months, more than reach as many users as possible in many years.

Thus, we can deal with the following problem:

Definition 4 (Influence Maximization problem) *Given a sharing knowledge graph* KSN *G, a set of users U, k a desired number of influential people and a time-horizon T, the influence maximization problem consists in identifying a minimal subset of users that allows to maximize the spread of influence in G within T.*

In this formulation of the problem the aim is the identification of the subset of most influential customers in the KSN graph. Thus, this formulation of the problem led us to ask who are the most influential people and how we can identify them. To deal with this problem our idea is to leverage the hill climbing greedy strategy proposed by Kempe et al. [7] on the influence graph, exploiting the information obtained by influential paths. This strategy has the aim to maximize the influence function ϕ, defined as the number of activated user at the end of the process, leveraging the available influential paths.

Thus we propose the *Influential path greedy algorithm* (IPGA) 1. The idea behind our approach is to build iteratively the seed set S, choosing in each step a node that leads to obtain the largest increase in the estimated spread until the cardinality of S does not equals to k.

Algorithm 1 Influential path greedy algorithm

```
1: procedure Influential_Path_Greedy_Algorithm(G,K)
2:     S = ∅
3:     while |S| ≤ k do
4:         u ← arg max_{v∈V} φ
5:         S ← S ∪ u
6:     end while
7:     return S_0
8: end procedure
```

To better explain our idea we show how the described algorithm allows us to identify the two influential users in Figure 4. By using the proposed greedy strategy on the influential graph of Example 4 and choosing a value of k equals to 2, it is possible to note that the most influential users are *Picus* and *Giank*. In fact, as shown in Figure 4, other users in the network are more likely to be influenced by these two users.

5 Experimental analysis

In the following, we will describe the used methodology for evaluating the efficiency and the effectiveness of the introduced influence maximization algorithm over the examined network[4].

To perform the described evaluation, we used the Yelp dataset[5]: it is composed by information about 77 K local businesses in 10 cities across 4 countries (U.K., Germany, Canada, US). Moreover this dataset contains also 2.2 M of review made from 552 K of users connected through 3.5M of social edges. We enrich the examined dataset crawling different economic information from YELP website [6]; in particular, for each category, we use the unified id of each business object for extracting the related economic information that can regard the range price, menu price, ticket price by using the YELP API[7].

Experiments have been conducted exploiting ENEAGRID/CRESCO High Performance Computing infrastructure resources. ENEAGRID consists in computational and storage resources (more than 8000 cores and 1.5 PB of storage) located in 6 ENEA research centers interconnected by means of GARR network. Among the 6 sites, Portici Research Center is the most relevant one hosting the newest CRESCO [8] clusters [4].

To perform our experiments, we set up an environment for big data processing using Spark and Hadoop. We set up a cluster in a cloud-based environment consisting of 3 computing nodes with 4 cores 16GB RAM each one.

We show the evaluation made to suggest possible market strategy based on trend set analysis and user preferences.

Firstly, we measured the execution time of our approach of influence maximization varying the size of the related seed set k.

Successively, we evaluated the spread of influence with respect to a given ground-truth. The ground-truth is composed by a set of influential users using a strategy that takes in account how many relevant reviews have been made by each user. For this reason, we define the *popularity* value for each user, based on a combination of votes (useful, funny and cool votes) that users can assign to its reviews.

The trend of running time varying the cardinality k of seed set from 1 to 50 is shown in figure 5.

We use a recall based measure to evaluate the effectiveness of our approach with respect to the generated ground truth:

$$R = \frac{|\hat{U} \cap \tilde{U}|}{|\hat{U}|}$$

[4] For the provided analysis, we set the value of normalizing factor α equals to 0.5
[5] https://www.yelp.com/dataset_challenge
[6] www.yelp.com
[7] https://www.yelp.com/developers/documentation/v2/overview
[8] http://www.cresco.enea.it

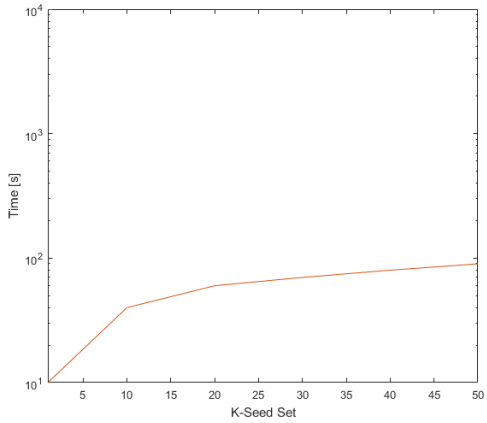

Fig. 5: Efficiency of the IPGA varying k

where $\hat{U}$ corresponds to the set of influentials in the ground-truth and $\tilde{U}$ is the set of influentials computed by the influence maximization algorithm.

The obtained results of recall have been shown in Table 1, computed on both the following graphs and considering the seed set cardinality k equals to 50, following the parameters settings in [11]: the proposed IG and the graph, called FG, generated considering only friendship relationships.

	IPGA (IG)	IPGA (FG)
Recall	78%	67%

Table 1: Recall Evaluation

As we can see from the results shown in the table 1, our method provides better performance with respect to the approach based on friendship relationships because it allows to consider several data, including also financial information.

These types of data contribute to properly define the influence exerts from users on the others that corresponds to a combination of the relevance of reviews made by them and the price range of business objects that they attend. These types of information are useful in modern market in which the dynamic interactions among users can lead the business objects strategies. Thus, based on the data involved in the proposed model (financial information, geographical information and so on), each firm can decide what is the better strategies to spend the budget for its market campaign. Eventually, each company can choose among the following strategies in each time horizon: influence new users by providing them discounts or gifts, consolidate influenced users through specific offers for them, or cooperate with other business objects.

6 Conclusion and future works

In this paper we have presented a data model for social network in order to analyze the influence among users. This approach can be useful for several applications such as trend analysis, viral marketing, location mining, urban planning and so on. In particular, the influence maximization problem is surely useful for increasing the profits obtained by marketing plan. The identification of the most influential people, in fact, allows firms to better optimize the marketing plan, tailoring the action for each customer to his preferences.

References

1. A whopping 20% of yelp reviews are fake (2013)
2. Borgs, C., Brautbar, M., Chayes, J., Lucier, B.: Maximizing social influence in nearly optimal time. In: Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms. pp. 946–957. Society for Industrial and Applied Mathematics (2014)
3. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 57–66. ACM (2001)
4. G. Ponti et al.: The role of medium size facilities in the HPC ecosystem: the case of the new CRESCO4 cluster integrated in the ENEAGRID infrastructure. In: International Conference on High Performance Computing & Simulation, HPCS 2014, Bologna, Italy, 21-25 July, 2014. pp. 1030–1033 (2014)
5. Granovetter, M.: Threshold models of collective behavior. American journal of sociology 83(6), 1420–1443 (1978)
6. Jung, K., Heo, W., Chen, W.: Irie: Scalable and robust influence maximization in social networks. In: 2012 IEEE 12th International Conference on Data Mining. pp. 918–923. IEEE (2012)
7. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 137–146. ACM (2003)
8. Newman, M.E.J.: Networks : An Introduction (2010)
9. Shehnepoor, S., Salehi, M., Farahbakhsh, R., Crespi, N.: Netspam: A network-based spam detection framework for reviews in online social media. IEEE Transactions on Information Forensics and Security 12(7), 1585–1595 (July 2017)
10. Tang, Y., Shi, Y., Xiao, X.: Influence maximization in near-linear time: A martingale approach. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. pp. 1539–1554. ACM (2015)
11. Tang, Y., Xiao, X., Shi, Y.: Influence maximization: Near-optimal time complexity meets practical efficiency. In: Proceedings of the 2014 ACM SIGMOD international conference on Management of data. pp. 75–86. ACM (2014)
12. Wang, Y., Cong, G., Song, G., Xie, K.: Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1039–1048. ACM (2010)
13. Zhang, J.D., Chow, C.Y., Xu, J.: Enabling kernel-based attribute-aware matrix factorization for rating prediction. IEEE Transactions on Knowledge and Data Engineering (2016)

14. Zhu, W.Y., Peng, W.C., Chen, L.J., Zheng, K., Zhou, X.: Exploiting viral marketing for location promotion in location-based social networks. ACM Trans. Knowl. Discov. Data 11(2), 25:1–25:28 (Nov 2016), http://doi.acm.org/10.1145/3001938