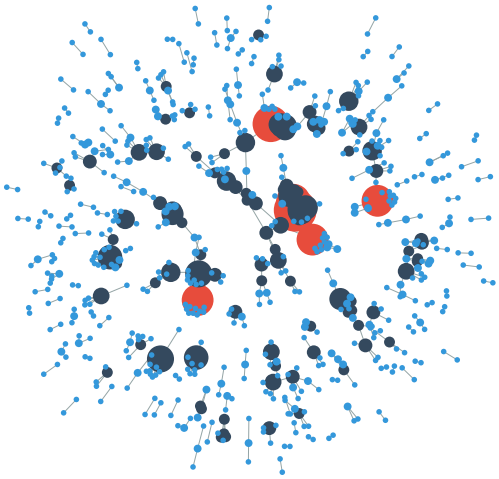
REVEALING INFORMATION NETWORKS

PH.D. THESIS

BOOKLET



Author
Róbert Pálovics
Institute for Computer Science and Control
Hungarian Academy of Sciences

Supervisor
András A. Benczúr
Institute for Computer Science and Control
Hungarian Academy of Sciences

BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS

2017

Recently, online social network sites such as Twitter and Facebook have emerged under the era of the Web 2.0. They catalyze global communication by allowing their users to immediately share information with each other. The capacity to store massive amounts of data has resulted in large datasets behind these services. Online social media sites have rich information about their users that may include their timely activity, social contacts, group behavior, or location information. The collected data gives an exceptional chance to analyze the behavior of the individual within a social community, and attracted several researchers from the domains of mathematics, physics, sociology, computer science and biology. The term "*computational social science* is occurring ... with an unprecedented breadth and depth and scale" [7].

The study of *complex networks* is an interdisciplinary field that recently received high attention. Several network science papers investigate the society by modeling it as a large graph: nodes correspond to users and edges represent relations, e.g. friendship or information sharing. Some of the main research directions are (1) characterizing the structure of online social networks, (2) identifying communities by finding highly connected components, (3) predicting social links that are likely to appear in the near future, (4) understanding how information spreads through social contacts in the global social network.

While the WWW is an exceptional online laboratory for the scientists, the scale of the daily generated content on the web results in a daily challenge for the single user. "We are drowning in information but starved for knowledge" [9]. Key problem for the users of the WWW is to filter and find relevant articles, products, or shared content in the social media. Beside search engines, *recommender systems* may aid the users by collecting, organizing, and ranking the information in online services. Recommenders become an industry standard since The Netflix Prize Competition [2] and are applied in a wide variety of domains including products, articles, news, movies, music, or books. In order to retrieve a ranked top list of relevant items for the user, recommenders may utilize user profile, metadata, browsing history, or context.

The main goal of our research is to analyze and understand the patterns of information flow in the global social network. Specifically, we intend to answer the following questions:

- How do information networks grow? What effects drive the growth of information networks?
- What are the indicators of social influence? Is it possible to use the effect of influence in recommendation systems?
- How can we mine the patterns of information flow at a global scale in order to utilize it in recommenders?

The results in this thesis attempt to answer these questions. Our findings are related to the research of *complex networks*, and *recommender systems*. In our work we analyze logs of social media sites containing timeline information about their users. In our experiments we include data from systems that contain social network and geolocation information.

Our contributions

Next we explain our main results one-by-one. For each topic, we list our main contributions and the original source of publication.

1 Raising graphs from randomness

In the research of complex networks, numerous results focus on one single statistic, the degree distribution. Barabási et al. [1] claim that the degree distribution of several real-world networks is a power-law function, i.e. the probability that a given non-zero degree node has degree k in the network, $p(d(i) = k)$ is

$$p(d(i) = k) = Ck^{-\alpha}, \qquad \alpha > 2.$$

Several previously proposed graph models result in networks with power-law degree distributions.

Another heavily investigated, yet simple statistic is the average degree. In growing systems like social networks, the average degree naturally increases with the network growth. Some state that the average degree is a *power-law* function of the number of nodes n,

$$\overline{d}(n) = cn^b.$$

The effect has been named *accelerated growth* by Dorogovtsev et al. [3], and *densification law* by Leskovec et al. [8].

We study the growth of information networks by considering processes where each node and edge is added to the network only once, and no node or edge is deleted from the network. We measure the average degree and the evolution of the degree distribution in these *growing networks*. Our experiments are based on three Twitter at-mention networks and three more from the Koblenz Network Collection [5], a publicly available network database.

Our results:

Thesis 1: *We propose a new model for the growth of information networks. We measured in real growing information networks that the average degree increases as $a + cn^b$ while the exponent of the power-law degree distribution decreases down to 2. Our preferential attachment and exponential growth based model is capable of reproducing these effects.*

- The average degree grows as $a + cn^b$, where n is the number of nodes in the network. We emphasize the importance of the constant a in the average degree formula. The constant was considered negligible in the experiments of Leskovec et al. [8]. In our results, however, the constant helps to capture the mixture of edges that appear at random vs. as a result of common interest, and fit to the actual measurements.
- The degree distribution of the network remains power-law during the growth, but the exponent of the power-law decreases. Note that one of most well-known models for growing networks is the Barabási-Albert model [1]. In case of this model the degree distribution exponent stays very close to constant as seen in Figure 1 left. In contrast in our measurements the degree distribution log-log plot lines of real networks get flattened (see Fig. 1 right).
- We present a series of measurements, where we precisely compute the above two statistics, and connect the growth of the average degree to the decay of the power-law exponent.
- We propose a model related to preferential attachment and exponential growth capable of reproducing both increasing average degree and the decreasing power law exponent. In our growing network model we add at each time step: (*i*) *random* new edges that connect two new nodes in the network, (*ii*) and *homophily* edges between already existing nodes in the network. More specifically, at time t, when the number of nodes is $n(t)$ in the network:

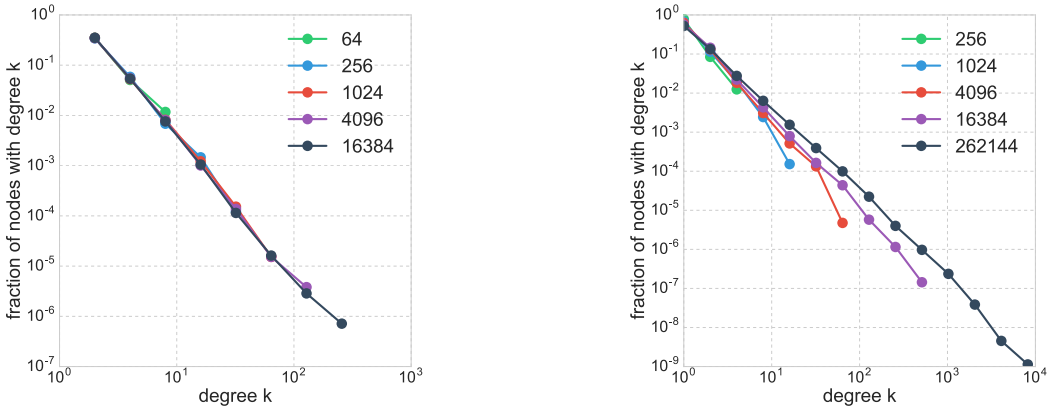


Figure 1: Degree distribution snapshots of growing networks at different sizes (number of nodes) indicated in the legend. **Left:** The Barabási-Albert model yields fixed exponent. **Right:** The Occupy Twitter mention data set with flattening slope as the network grows.

- For some constant r, $r \cdot n(t)$ new *random edges* appear that indicate the random growth of the network.
- Each node i selects other nodes to connect with *homophily edges* randomly. The expected number of new homophily connections created by node i is $s \cdot d_h(i)$, where $d_h(i)$ is the number of homophily edges already connected to node i, i.e. the homophily degree of node i. For a given new connection of node i, the target node is selected by preferential attachment. In other words, the probability of selection for node j as a new neighbor of i is the degree of j $d(j)$.

The main difference of our model compared to earlier models can be summarized in three points.

- The power law exponent, as in all our real networks, is greater than 2, this could not be modeled in [8].
- Our model explains the initial behavior of the degrees as a natural mixture of influence and preferential attachment edges, and also predicts correctly the ratio of these edges.
- Our model generates both increasing average degree and decreasing power law exponent.

As a general overview of the possible models based on our observations, networks start to grow at random, like an Erdős-Rényi graph. Then certain rules such as preferential attachment [1] intensifies during the growth process, and causes scale-free degree distribution with a decreasing exponent. The stronger the rule is, the closer the exponent of the degree distribution gets down to two in a more coherent network. As the degree distribution log-log plot flattens, the chance for very high degree nodes in a strongly skewed distribution increases acting as the main organizer of the network structure.

Our finding appeared as

I. Róbert Pálovics and András A Benczúr. Raising graphs from randomness to reveal information networks. In *Proceedings of WSDM 2017*, 2017.

2 The online ranking prediction problem

Next we turn to the research of recommender systems that has become popular since the Netflix prize [2]. Recommender systems serve to predict preferences of users on items. They offer relevant items for users in systems where the available set of items is too large. Examples for recommender systems are: (1) music recommender algorithms in music streaming services, e.g. Spotify, (2) recommendation of movies in online movie catalogs, e.g. Netflix, (3) recommendation of items in online webshops, e.g. Amazon. Recommenders are information filtering algorithms that select for the user relevant items that she may consider.

In 2009, the Netflix Prize [2] resulted in an increased popularity of the research of recommenders in computer science. The contest was defined as a *batch rating prediction* task, with one part of the data used for model training, and the other for evaluation. However, usually recommender systems should present a ranked top list of relevant items for the user. Moreover, users request one or a few items at a time and get exposed to new information that may change their needs and taste when they return to the service next time. Furthermore, recommenders often utilize the context information of the user that can be non-stationary, like geolocation information.

In a real application, *top item recommendation by online learning* is hence apparently more relevant than batch rating prediction. However, this task received much less attention. In our work we consider top recommendation in highly non-stationary environments. Our goal is to promptly update recommender models after user interactions by online learning methods. Our contributions:

Thesis 2: *We propose the online ranking prediction problem for recommender systems that better approximates the current needs of online services. Furthermore, we define the online stochastic gradient based matrix factorization algorithm as a robust baseline for this temporal ranking prediction task.*

- We formalize the problem of recommending in highly non-stationary environments by defining the *online ranking prediction problem*. In this setting, the model should retrieve a top-k recommendation list for each event in the time series by learning from the past events in the data. In other words the personalized user top-k recommendations are continuously updated over time. In an online setting,
  - we query the recommender for a top-k recommendation for the active user,
  - we evaluate the list in question against the relevant item that the user interacted with,
  - we allow the recommender to train on the revealed user-item interaction.
- We propose matrix factorization [4] by online stochastic gradient as a time-aware baseline algorithm. Note that this is particularly relevant as factorization models are considered as one of the strongest baselines in stationary environments. As originally designed, stochastic gradient descent methods may iterate several times over the training set until convergence. In an real-time recommendation task, however, the model needs to be re-trained after each new event and hence re-iterations over the earlier parts of the data may be computationally infeasible. We implement an online matrix factorization algorithm by allowing a *single* iteration over the training data only, and in this single iteration, we process the events in *temporal* order.
- We introduce personalized algorithms to combine different recommender methods online, and apply them on our new context-aware methods presented in the next sections.

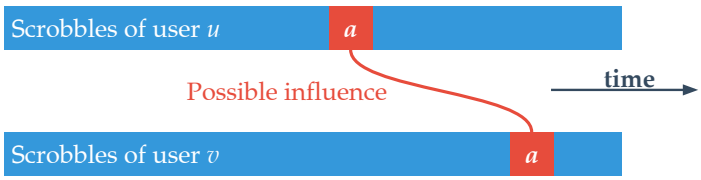


Figure 2: Possible influence between two users u and v that both listened to the same artist a.

Note that our results are presented in our papers on temporal context-aware recommendation methods. Hence the corresponding papers are listed in the next sections.

3 The effect of social influence

Next we detect and investigate the effect of social influence. Our experiments are conducted over data from Last.fm, a social media site that tracks the music listening behavior of its users. In Last.fm terminology, a music listening event is called a "scrobble" in their database. Key property of the dataset is that no exact information is available on social influence. In our work we intend to detect influence by using the fact that it results in correlation between the behavior of friends. As indicated in Figure 2, we detect events when two friends in a social network scrobble the first time the same artist after each other. These events may be the result of social influence between friends: they may induce each other to adopt a new behavior, i.e. listen to a new artist. However, detecting influence is a hard task in general, since other social effects, most notably homophily, may result in similar correlation effects. Our main findings:

Thesis 3: *We detect social influence in systems where there is no explicit information on influence between individuals in a community. We utilize our framework that predicts the probability of influence between friends and create an influence based recommender system.*

- We present a method to measure social influence in Last.fm. Note that the data set has a very general structure, and we expect that our methods may work on data from other domains. We give the theoretical background of our results by developing a model for the probability of influence between two friends in the social network. Based on our measurements, we estimate the probability of influence as the cause of a given event.
- We give a framework of a system that distills social influence to recommend new music for the users. We examine our new recommender method in experiments defined by the online ranking prediction problem.
- We give a more sophisticated version of our recommender by modeling correlation effects with context-aware matrix factorization.
- We successfully combine the variants of our influence recommender with online matrix factorization, a strong baseline that exploits homophily effects and bursty behavior.

This chapter is a summary of our work presented in a series of papers,

II. Róbert Pálovics and András A Benczúr. Temporal influence over the Last.fm social network. In *Proceedings of IEEE/ACM ASONAM 2013*, pages 486–493. ACM, 2013,

III. Róbert Pálovics and András A Benczúr. Temporal influence over the Last.fm social network. *Social Network Analysis and Mining*, 5(1):1–12, 2015,

IV. Róbert Pálovics, András A Benczúr, Levente Kocsis, Tamás Kiss, and Erzsébet Frigó. Exploiting temporal influence in online recommendation. In *Proceedings of RecSys 2014*, pages 273–280. ACM, 2014.

4 Hierarchical models for geolocation data

Finally we analyze social networks with geolocation information. We design position based recommender methods that, in addition to user preferences, also learn item locality. Since items may have a very strong time dependence at a location, we consider methods of recommendation by online machine learning. We develop a context-aware recommender system for the online ranking prediction problem that uses the updated geoinfo of the users. For our experiments, we construct data based on Twitter, a service that can be considered as a mix of a social network and news media [6], and in addition, an information system with geographical information. We investigate the problem of recommending Twitter hashtags for users, based on the temporal geolocation information of both the users and the hashtags.

Our results:

Thesis 4: *We propose location based hierarchical recommendation models that are capable of detecting and predicting the diffusion of trends in social media. Our results are applicable in scenarios when the data is too sparse for factorization based recommendation.*

- We define recency and popularity based recommender algorithms that can be applied in sparse datasets.
- We inject our algorithms into a location-aware recommender structure, where we define local models. Our key idea is that we organize the local models to a hierarchical structure based on geolocation, as seen in Figure 3. For example, we recommend relevant items based on the recent events in the user's current neighborhood, city, country, and continent.
- As a result, our models are capable of detecting the diffusion of trends in social media at a global scale.
- We learn the combination weights of the local models by the online stochastic gradient algorithm.
- We define for each hierarchical model its baseline version and analyze our methods in-depth.

Our results are published in

V. Róbert Pálovics, Péter Szalai, Júlia Pap, Erzsébet Frigó, Levente Kocsis, and András A Benczúr. Location-aware online learning for top-k recommendation. *Pervasive and Mobile Computing*, 2016.

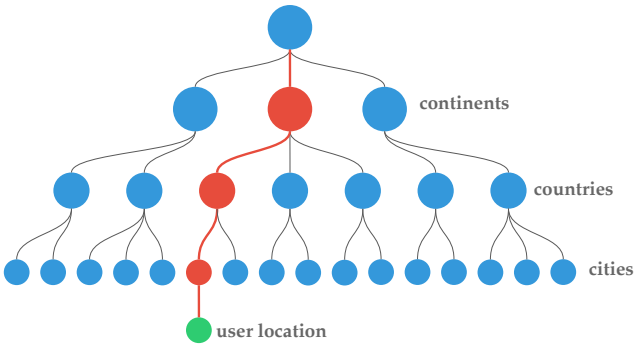


Figure 3: Hierarchical model for geolocation based recommendation.

Related publications

The corresponding publications are summarized in the following list:

I. Róbert Pálovics and András A Benczúr. Raising graphs from randomness to reveal information networks. In *Proceedings of WSDM 2017*, 2017

II. Róbert Pálovics and András A Benczúr. Temporal influence over the Last.fm social network. In *Proceedings of IEEE/ACM ASONAM 2013*, pages 486–493. ACM, 2013

III. Róbert Pálovics and András A Benczúr. Temporal influence over the Last.fm social network. *Social Network Analysis and Mining*, 5(1):1–12, 2015

IV. Róbert Pálovics, András A Benczúr, Levente Kocsis, Tamás Kiss, and Erzsébet Frigó. Exploiting temporal influence in online recommendation. In *Proceedings of RecSys 2014*, pages 273–280. ACM, 2014

V. Róbert Pálovics, Péter Szalai, Júlia Pap, Erzsébet Frigó, Levente Kocsis, and András A Benczúr. Location-aware online learning for top-k recommendation. *Pervasive and Mobile Computing*, 2016

Bibliography

[1] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[2] James Bennett and Stan Lanning. The Netflix prize. In *KDD Cup and Workshop in conjunction with KDD 2007*, 2007.

[3] Sergey N Dorogovtsev and JFF Mendes. Accelerated growth of networks. *arXiv preprint cond-mat/0204102*, 2002.

[4] Yehuda Koren, Robert Bell, Chris Volinsky, et al. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[5] Jérôme Kunegis. Konect: the Koblenz network collection. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1343–1350. International World Wide Web Conferences Steering Committee, 2013.

[6] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[7] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[8] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007.

[9] John Naisbitt and J Cracknell. Megatrends: Ten new directions transforming our lives. Technical report, Warner Books New York, 1984.

[10] Róbert Pálovics and András A Benczúr. Temporal influence over the Last.fm social network. In *Proceedings of IEEE/ACM ASONAM 2013*, pages 486–493. ACM, 2013.

[11] Róbert Pálovics and András A Benczúr. Temporal influence over the Last.fm social network. *Social Network Analysis and Mining*, 5(1):1–12, 2015.

[12] Róbert Pálovics and András A Benczúr. Raising graphs from randomness to reveal information networks. In *Proceedings of WSDM 2017*, 2017.

[13] Róbert Pálovics, András A Benczúr, Levente Kocsis, Tamás Kiss, and Erzsébet Frigó. Exploiting temporal influence in online recommendation. In *Proceedings of RecSys 2014*, pages 273–280. ACM, 2014.

[14] Róbert Pálovics, Péter Szalai, Júlia Pap, Erzsébet Frigó, Levente Kocsis, and András A Benczúr. Location-aware online learning for top-k recommendation. *Pervasive and Mobile Computing*, 2016.