Exploring social network effects on popularity biases in recommender systems

Rocío Cañamares and Pablo Castells
Universidad Autónoma de Madrid, Escuela Politécnica Superior, Departamento de Ingeniería Informática
{rocio.canamares,pablo.castells}@uam.es

ABSTRACT

Recommending items ranked by popularity has been found to be a fairly competitive approach in the top-N recommendation task. In this paper we explore whether popularity should always be expected to be an effective approach for recommendation, and what causes, factors and conditions determine and explain such effectiveness. We focus on two fundamental potential sources of biases in rating data which determine the answer to these questions: item discovery by users, and rating decision. We research the role of social communication as a major source of item discovery biases (and therefore rating biases). We undertake the study by defining a probabilistic model of such factors, and running simulations where we analyze the relationships between the effectiveness of popularity and different configurations of social behavior.

Keywords

Popularity, social networks, evaluation, viral propagation.

1. INTRODUCTION

Recommending items ranked by popularity has been found to be a fairly competitive approach in the top-N recommendation task [5]. It may be to some initial surprise that a trivial and non-personalized recommendation method can be this effective, somewhat contradicting the implicit intuition underlying the recommender systems field that personalized recommendations should have the potential to maximize overall user satisfaction, by achieving an optimal fit of users' needs on an individual basis, as opposed to a one-size-fits-all approach.

Some authors have analyzed this issue recently [8,11,12,13,14], and have proposed specific techniques to consider the biases in the distribution of missing ratings, in both the recommendation algorithms and the evaluation methodology and metrics. The question has also been addressed from the perspective of the actual utility of recommendation: recommending popular items has the obvious shortcoming of a lack of surprise for the user, approximating (by definition) the worst possible results in terms of the novelty dimension [4]. Despite this obvious shortcoming, popular recommendations appear to be reasonably effective in practice (e.g. as a fallback option), item popularity is actually an (intentional or accidental) ingredient of many state of the art recommendation algorithms, and commercial applications seem to be using it among other signals in recommendation functionalities. However, we barely find in the literature a clear analysis of the causes and characteristics of the popularity biases, and the relationship between the popularity distribution and the potential consequences in the performance and evaluation of recommendation algorithms.

It is natural to wonder a) whether popularity should always be expected to be an effective approach (or partial signal) for recommendation, b) what causes, factors and conditions determine and explain such effectiveness, and c) whether the apparent effectiveness actually reflects true effectiveness, or is the result of a distortion of some sort in the evaluation methodologies. We address such questions in this paper.

Popularity-based recommendation exploits biases in the distribution of available observed ratings among items –or equivalently, of the distribution of missing ratings. Thus studying the properties of popularity is essentially the same as studying the characteristics of rating distributions, and their biases. In unbiased situations (where ratings are uniformly distributed), popularity is equivalent to random recommendation and makes no particular sense as a recommendation strategy. Popularity therefore makes sense when rating data is biased or, in other words, missing not at random [7,11].

In this paper we focus on two fundamental potential sources of biases in rating data: item discovery biases, and rating decision biases. The latter refers to the factors that determine whether or not a user decides to rate an item he has interacted with; for instance, in many cases users may be typically more prone to rate items they have liked than items they have not liked. The former refers to the fact that in order to be rated by a user, the user needs first to become aware that the item exists. Biases in item discovery distribution then naturally result in biases in the items that ultimately get more ratings or less.

Discovery biases are determined by the sources by which users discover items. People get to know items through a variety of channels such as direct user searches, advertisement from providers, random encounter, suggestions from a recommender system, etc. Beyond this and foremost, our social environment is a key source of information and discovery for which people have a particular reliance and trust compared to other channels. The perspective of the role word of mouth has in the distribution of ratings connects the problem at hand to an issue of network propagation: the items that propagate faster and farther in the social network will tend to get more ratings.

Propagation phenomena have been extensively studied in the area of complex networks, and social networks in particular (for diseases, rumors, viral effects, etc.) [1,6,9,10], but with scarce exceptions [3] the connection to biases in user rating distribution have been barely examined before. Yet we find that network effects can be a major explanatory factor for recommendation data biases and popularity effects.

In this paper we address this perspective. We posit in particular the following potential key factors in creating popularity biases, determining whether popularity becomes or not a good strategy to achieve recommendation effectiveness:

- User behavior in their communication with peers, in particular the biases towards positive or negative experiences when sharing one's experiences with others, and the overall frequency with which users intercommunicate in social networks.

Proceedings of the 6th Workshop on Recommender Systems and the Social Web (RSWeb 2014), collocated with ACM RecSys 2014, 10/06/2014, Foster City, CA, USA. Copyright held by the authors.

- User behavior in rating decisions, in particular, biases towards rating positive or negative experiences.
- Social network structure, in particular link density and clustering structure.

We undertake this study by representing the involved factors in a probabilistic model defined by random variables subject to interdependent distributions. Based on the model, the problem can be approached, complementarily, by a formal analysis, or by empirical observation through simulations. In this paper we pursue the latter path. We identify the key variables, parameters and dependencies describing the factors we aim to focus on and we explore, through simulation based on the proposed model, the resulting effects on the effectiveness of popularity, aiming to identify different situations and uncover potential explanations thereof.

2. A SOCIAL RATING GENERATION MODEL

We start our analysis by formalizing the fundamental actions, events and variables involved in the rating generation process, upon which we will formally identify and formalize the key factors for the phenomena we aim to observe (user behavior trends and related network processes), and their relations to resulting effects (data biases and effectiveness variations in popularity-based recommendation), in the form of probabilistic dependencies and model parameters.

In order to generate input data for an item, a user needs to become aware that the item exists, decide to interact with the item, and then decide to rate it. Popularity biases in recommender system input data can be therefore related to two main factors: a) biases in the items that users discover: some items become known to many more users than others; and b) biases in the items that users decide to rate (or consume or interact with): once a user experiences an item, there may be some systematic reason why users decide to rate certain items and not others.

The primary necessary steps by which a rating value is generated can be thus identified as follows:

1. A user discovers an item, i.e. he becomes aware that the item exists.
2. The user decides to interact with (consume, click, play, etc.) the item.
3. The user decides to rate the item.

Moreover, in a social environment, we consider an additional relevant action by users on items:

4. The user shares with some of his friends his experience with the item. This brings to step 1 (discovery) each person informed by the user about the item.

The distinction between steps 2 and 3 is not a clear cut or simply inexistent in common applications, where users do not enter explicit ratings, and user-item interaction data are used as input instead by recommendation algorithms; for this reason and the sake of simplicity we shall ignore the difference in our model.

These steps thus create a cycle by which users become aware of items or, from the item perspective, items progressively traverse the social network of users, becoming known to the users they come across, and becoming rated by some of them. How far and what regions an item reaches in the network depends on the intrinsic communication patterns of users in the network, the dependence of the latter on characteristics of the items, and the shape and connectivity of the network, which is known to affect the development of network propagation phenomena [6].

2.1 Random variables and parameters

We formally model the described process in terms of a set of binary random variables defined upon different sample spaces combining the set $\mathcal{U}$ of all users, the set $\mathcal{I}$ of all items, and the set $\mathcal{T}$ of all time points we may consider in the model, as follows:

- **Rating:** $rated: \mathcal{I} \times \mathcal{U} \times \mathcal{T} \to \{0,1\}$ takes value 1 for a sampled element (i,u,t) if user $u \in \mathcal{U}$ has rated item $i \in \mathcal{I}$ by or before time $t \in \mathcal{T}$, and 0 otherwise.
- **Relevance:** $relevant: \mathcal{I} \times \mathcal{U} \to \{0,1\}$ takes value 1 if the sampled user likes the sampled item, and 0 otherwise. Notice that this variable is not observed unless it becomes visible to the system when the user rates the item. Note also that we assume as a simplification that relevance is a static condition and does not change with time or context.
- **Discovery:** $seen: \mathcal{I} \times \mathcal{U} \times \mathcal{T} \to \{0,1\}$ is 1 if the user is aware the item exists by or before the time point at hand, and 0 otherwise. The same as relevance, this variable is not observed unless a user rates an item, in which case we know he must have seen it to begin with.

  As mentioned before, we ignore the distinction between knowing an item exists (e.g. we have seen a movie title on a billboard), and actually experiencing the item (e.g. we actually watch the movie). The difference can be worth being considered as it involves a decision on the part of the user, but it is not required for the focus of our present analysis.
- **Communication:** $tell: \mathcal{U} \times \mathcal{I} \times \mathcal{U} \times \mathcal{T} \to \{0,1\}$ is 1 if a user tells a given friend about a given item at a given time when both friends talk to each other, and 0 otherwise.

  We consider that users only share experiences with people they are connected with in a social network. This does not involve any loss of generality, as we do not make any assumption on the nature of the network at this point. The simplifying restriction will be made in our experiments, where we will use or simulate specific social network structures and assume (as a simplification) they embody full knowledge of all connections between users.

The key distributions and dependencies which capture the relevant factors in the behavior of the defined model can be expressed in terms of conditional probabilities. We would mainly foresee two such key dependencies: the propensity of users to rate items they like vs. items they do not like, and their inclination to share positive vs. negative experiences. This can be expressed by four conditional distributions:

$$p(tell|seen, relevant, u, i, v, t)$$
$$p(tell|seen, \neg relevant, u, i, v, t)$$
$$p(rated|seen, relevant, i, u, t)$$
$$p(rated|seen, \neg relevant, i, u, t)$$

If we make the simplifying assumption that the decision to rate and share mainly depends on the relevance of the item, and we ignore for a moment the differences between users in this respect, as well as the possible variations in user behavior over time, we may consider the approximation $p(tell|seen, relevant, i, u, t) \sim p(tell|seen, relevant)$ –and the same for the other four distributions– in such a way that we have four conditional probabilities defining two behavioral dimensions, which may act as configuration parameters of the model:

- **Communication/relevance bias:** $p(tell|seen, relevant)$ and $p(tell|seen, \neg relevant)$.
- **Rating/relevance bias:** $p(rated|seen, relevant)$ and $p(rated|seen, \neg relevant)$.

We are interested in studying how these parameters affect the effectiveness of popularity-based recommendation. For that purpose, we shall simulate an environment the dynamics of which are based on the proposed model, run recommendations in that environment using the generated ratings, and measure their effectiveness according to the simulated data. The model defines a few rules that changes of state in the environment should be governed by, but in order to simulate the environment dynamics we need to define a set of triggering actions and events, and the order in which they take place.

2.2 Model dynamics

We consider the following simplified scenario that represents how items may become known to users and eventually rated. We have a population of users and a set of items. Based on the model defined in the previous subsection, user-item pairs undergo a sequence of states, from unknown to discovered to rated, in this order, where the two latter states may or may not be ever reached. Items are discovered by users through friends: at certain points in time, users choose a friend and an item they have discovered, and decide whether or not to tell the friend about the item. If communication takes place, the friend discovers the item the user talks about (if he had not discovered the item already). Communication takes place as a dialog, which means that the friend will in turn choose some discovered item and (under the same relevance-based communication probability pattern) talk back about it to the first user, who will then discover this item. In our chosen configuration, users talk about an item on their own initiative only once at most, but they can talk about it any number of times when asked.

Users thus discover items by communication through the social network. However, initially all items are unknown to all users. In order to bootstrap the system, we may either define an initial state where an arbitrary set of user-item pairs are in the discovered state (e.g. n random users have discovered each item), or we include an additional discovery source, extrinsic to the social network, through which items may also become known. In our current implementation we choose the second option. The source may represent e.g. catalog browsing and searching, item advertisement, etc., and can be implemented as random sampling (as slow and infrequent as we would desire with respect to the overall simulation time flow) of user-item pairs for discovery, or biased sampling by some arbitrary distribution, or even a recommender system. In our case we choose random sampling by a ratio of 0.1% of the simulation time step (that is, on average every 1 out of 1,000 simulation steps, users discover an item at random with replacement –i.e. we do not force the sampled item to be unknown and it may have been discovered already).

The decision to rate items and to share the experience with friends can be required of the simulated users in different ways and order. As a simplification, the decision to rate an item or not is made at the time when the user discovers the item. If the user does not rate it, the decision is not reconsidered anymore. Regarding communication, in our chosen configuration each user is given a chance to talk about an item to a friend once every simulation time unit –or inversely, the time unit is defined as an iteration where every user is given the chance to speak to a friend. The item is chosen uniformly at random (without replacement if the user took the initiative) among the ones the user has discovered, and the friend is sampled uniformly at random (with replacement) from all the user's social contacts.

The rating and sharing decisions, when the user is faced to them as explained in the previous paragraph, are taken based on the probabilistic model described in the previous subsection. That is, when a user discovers an item, he will rate it with probability $p(rated|seen, relevant)$ if the user likes the item, and with probability $p(rated|seen, \neg relevant)$ if he does not. Analogously, the decision to talk or not to a friend about an item is taken (once the friend and the item have been sampled) according to $p(tell|seen, relevant)$ if the user likes the item, and $p(tell|seen, \neg relevant)$ if he does not.

In order to carry out the above simulated actions, it is apparent that we should know whether a given user likes a given item at the time when this determines the probabilities of the user's decisions. Relevance is in general an unobserved variable for the system (until a user rates an item) and for the user himself (until he discovers an item). We deal with this lack of observation by simulating relevance knowledge as a certain user-item relevance distribution. This knowledge will remain hidden to the system (in particular to the recommender systems we will run in our experiments), but will be made "visible" to a) the simulated users when they discover an item, and b) the computation of recommendation effectiveness metrics, as we will explain shortly.

Our model does not make any assumption about the relevance distribution, but in our experiments, we assume the number of users who like an item (which is equal to $p(relevant|i)$ multiplied by the number of users) has a long-tail distribution shape. This is an arbitrary decision in our work at this point, which could be contrasted by means of a poll of some kind in a real setting. It does not seem to be a critical aspect of the model in our simulations though. In order to obtain the long tail shape we use an adjusted power law defined by $p_\alpha(relevant|i_k) = c_1 + \beta(k + c_2)^{-\alpha}$ for the k-th most liked item. The parameter $\alpha \in [0, \infty)$ defines the steepness of the relevance distribution, where $\alpha = 0$ gives a uniform distribution. We adjust the remaining parameters c_1, c_2 and β in such a way that –we omit the details– the distribution adheres to a given prior $p(relevant)$, and the extremes of the curve for the most and least liked users (i_1 and $i_{|\mathcal{I}|}$) behave as one would expect, that is: $\lim_{\alpha \to \infty} p_\alpha(relevant|i_1) = 1$, $\lim_{\alpha \to \infty} p_\alpha\left(relevant\middle|i_{|\mathcal{I}|}\right) = 0$, $\lim_{\alpha \to 0} p_\alpha(relevant|i_1) = \lim_{\alpha \to 0} p_\alpha\left(relevant\middle|i_{|\mathcal{I}|}\right) = p(relevant)$. Figure 1 shows the shape of the curve for $\alpha = 1$ and $p(relevant) = 0.2$ with 3700 items.

Given a distribution thus defined, we generate the sequence of the number of users who like each item, and we assign this number randomly to the items. Then for each item, we assign the corresponding number of users liking the item by randomly sampling the users. We thus create a scenario where each user likes a set of items beforehand, although he does not know he likes an item until he discovers it. If the user rates the item, the system will also know whether or not the user likes it: if he does, the rating value will be

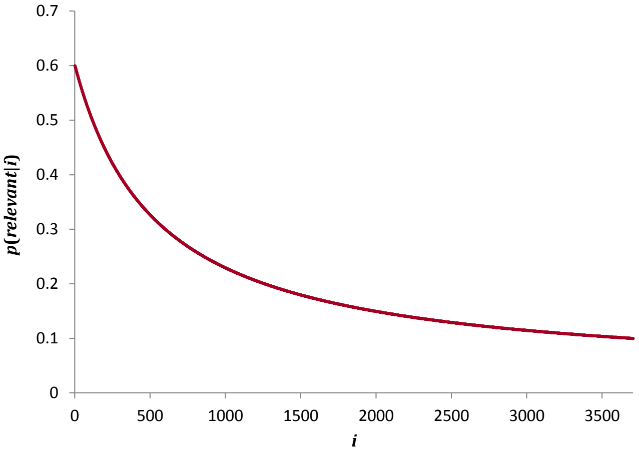

Figure 1. Simulated relevance distribution for $\boldsymbol{\alpha = 1}$ with prior $\boldsymbol{p(relevant) = 0.2}$ for $\boldsymbol{3700}$ items. Items are sorted from most to least liked in the $\boldsymbol{x}$ axis, and the line represents the ratio of users who like each item.

positive, and negative otherwise. We thus consider as a simplification that relevance is an immutable condition which does not change with time nor context.

3. ITEM POPULARITY RECOMMENDATION AND ITS EFFECTIVENESS

Before we get into the empirical analysis of popularity effectiveness, we cast precise definitions of popularity and the metrics to assess its effectiveness in recommendation. In the usual definition of popularity-based recommendation one finds in the literature, items are ranked by their total number of ratings, regardless of whether the rating values express a positive or negative preference [5]. Given the role of relevance we are analyzing in our model, we find it worthy to also consider a variant of popularity recommendation where only positive ratings are considered. We therefore study both variants in our experiments, the scoring functions of which are defined as:

Simple popularity: $s(u,i) = |\{v \in \mathcal{U} | r(v,i) \neq \emptyset\}|$

Relevant popularity: $s(u,i) = |\{v \in \mathcal{U} | r(v,i) \geq \tau\}|$

where $r(v,i)$ is the rating assigned to i by v, and τ is the threshold value at or above which the rating expresses a positive preference.

As a metric of recommendation effectiveness we shall take precision $P@k$, which is a simple to compute and analyze yet representative metric. We shall distinguish between the observed precision P_{obs}, in which a recommended item is considered relevant if it has been rated by the target user with a positive value, and true precision P_{true}, in which we use all the simulated relevance knowledge, including that which has not become known to the system in the form of ratings:

$$P_{obs}@k = \operatorname*{avg}_{u \in \mathcal{U}} \left|\{i \in R_u^k | r(u,i) \geq \tau\}\right| / k$$

$$P_{true}@k = \operatorname*{avg}_{u \in \mathcal{U}} \left|\{i \in R_u^k | u \text{ likes } i\}\right| / k$$

where R_u^k is the set of top k items recommended to u. Naturally P_{obs} is what offline experimental evaluations commonly report. This will allow us to contrast precision as it is usually measured in offline recommender system experiments, to the true precision defined by the full underlying user preferences which have determined the generation of ratings and their distribution in our model.

4. SIMULATION-BASED EMPIRICAL OBSERVATIONS

The proposed model allows representing different user behavior patterns by different model configurations using different parameter settings. In order to analyze the effect that such configurations produce on the effectiveness of popularity-based recommendation, we implement a simulation framework that runs the model dynamics described in the previous section. The framework supports the integration of an arbitrary set of recommendation algorithms by just having them adhere to a simple abstract API. At each simulation time step, the framework generates a temporal split of rating data with a 0.5/0.5 ratio of training/test data, runs all the recommendation algorithms, and evaluates the observed and true precision for each of them. This way we can monitor how the performance of recommenders evolves along the simulation. For the focus of the present paper we only observe three recommenders: popularity, relevant popularity, and random.

The basic research questions we aim to shed light on in our experiments are the following:

RQ1. How does the observed and true precision of popularity-based recommendation depend on the users' social communication patterns, and in particular the bias towards sharing positive vs. negative experiences?

RQ2. How does the observed and true precision of popularity-based recommendation depend on the user bias towards rating liked vs. non-liked items?

RQ3. Can certain social network characteristics and effects (such as its topology and viral phenomena) alter the dependencies between user behavior and popularity precision?

RQ4. May the observed and true precision disagree in terms of how popularity compares to random recommendation, as a consequence of particular social behavior patterns?

4.1 Experimental setup

For most of the observations we report next, we use the social network data from Facebook made available by J. Leskovec [9], containing 88,234 social connections among 4,039 users. Taking inspiration on the order of scale of MovieLens 1M, we take a total set of 3,700 items. We simulate a relevance distribution with prior $p(relevant) = 0.2$ and steepness $\alpha = 1$, which results in the distribution shown earlier in Figure 1. For discovery bootstrapping, in addition to word of mouth, users discover an item at random 1 out of every 1,000 simulation steps. We run all simulations until 500,000 ratings have been generated, which is half the size of MovieLens 1M. The reason for not running the simulations longer is to avoid the distorting saturation effects that eventually start to appear as a consequence of the implicit closed world assumption involved in having a fixed set of users and items all along. E.g. in the limit all users end up consuming and/or sharing most items regardless of their preferences and behavior patterns, just by reason of exhaustion of any better remaining option. In the results we present here, we run each simulation only once, in order to show how the observed patterns can be perceived even without averaging random effects (having informally checked that the variation is moderate when averaging).

In order to study the effects of the relevance biases in sharing and rating user behavior, we shall fix certain parameters and observe the variation of popularity precision as we vary the others.

4.2 Communication/relevance bias

To isolate the effect of communication biases, we take a relevance-neutral rating behavior by $(rated|seen, relevant) = p(rated|seen, \neg relevant) = 1$, that is, users always rate all items they discover. Then, we shall vary $p(tell|seen, relevant)$, $p(tell|seen, \neg relevant)$ and the prior $p(tell|seen)$ as we explain next.

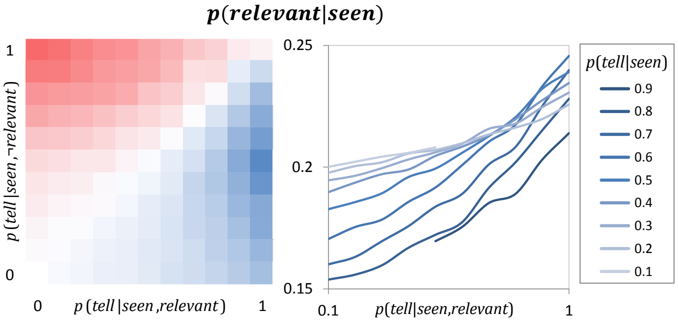

Figure 2. Discovery/relevance bias caused by communication/relevance biases. The color scale on the color map and the y axis on the graphic on represent $p(relevant|seen)$, blue being the maximum and red the minimum value. Note that the curve for $p(tell|seen) = 0.9$ has no values for $p(tell|seen, relevant) < 0.5$, as it is not possible to reach such a high prior with lower sharing probabilities on relevant items. Likewise, $p(tell|seen) = 0.1$ has no points for $p(tell|seen, relevant) > 0.5$.

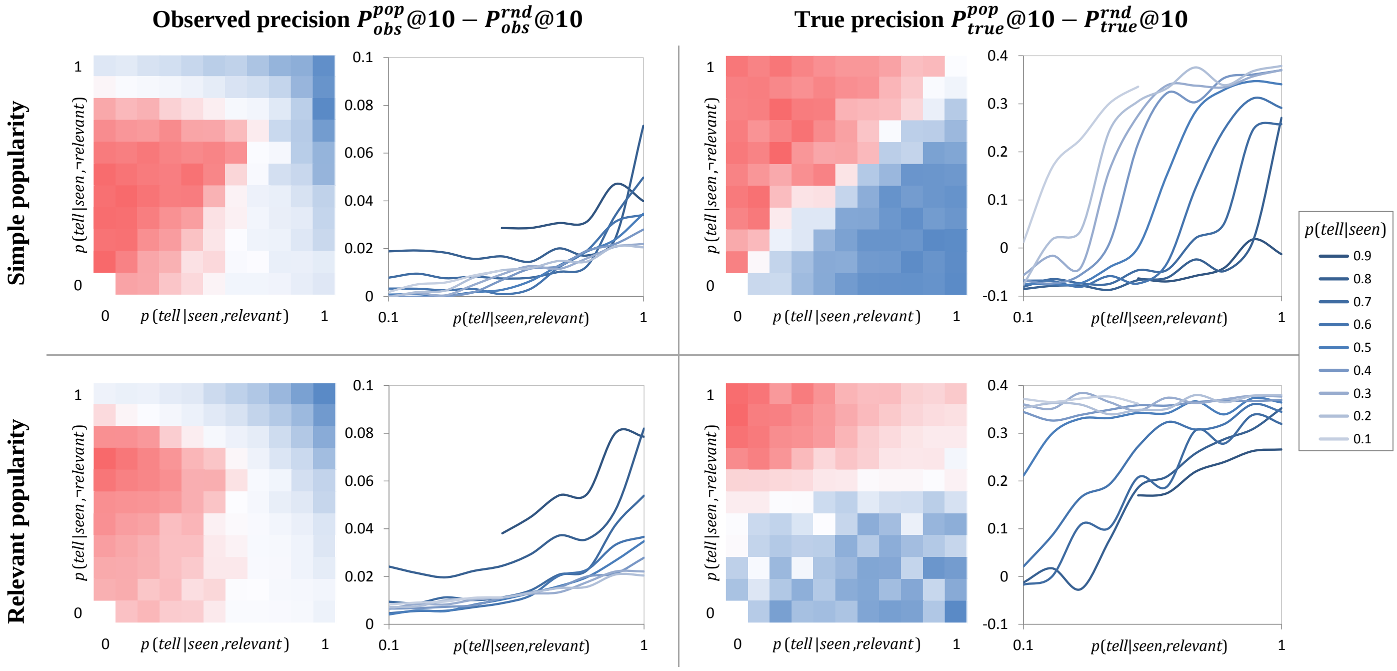

Figure 3. Effect of the communication/relevance biases on popularity-based recommendation precision. We show both the observed (left block) and true (right block) precision on two item popularity variants: ranking by the total number of ratings (simple popularity, top), and ranking by the number of positive ratings (relevant popularity, bottom). The precision values are actually shown as the difference to the precision of random recommendation. For each recmmender/metric combination, a color map shows the resulting precision values (blue being the maximum and red the minimum value) for the corresponding values of $p(tell|seen, relevant)$ and $p(tell|seen, \neg relevant)$. The $(0,0)$ cell is left empty as no communication takes place in such a setting. Next to each color map, a graphic shows a curve for each value of the prior $p(tell|seen)$ from 0.1 to 0.9 with $p(tell|seen, relevant)$ in the x axis, and the difference to random precision in the y axis. The same as in Figure 2 and for the same reason, the curves for $p(tell|seen) = 0.9$ have no values for $p(tell|seen, relevant) < 0.5$, and $p(tell|seen) = 0.1$ have no points for $p(tell|seen, relevant) > 0.5$.

4.2.1 Discovery/relevance bias

Before analyzing how the relevance-sharing biases affect popularity precision, we check a simple hypothesis that sharing biases result in approximately equivalent discovery biases. This is not as obvious as it might seem, as the fact that people speak about what they like (a relevance bias in communication) does not necessarily imply those who listen like it as well (a relevance bias in discovery).

Figure 2 shows how the communication/relevance bias results in a discovery bias in terms of the ratio of discovered items that are liked by the users who have discovered them, as expressed by $p(relevant|seen)$. The figure presents the results in two ways: in the color map on the left, we vary $p(tell|seen, relevant)$ and $p(tell|seen, \neg relevant)$ from 0 to 1 by increments of 0.1, and we show the resulting $p(relevant|seen)$ in a color scale, where blue is the maximum value and red is the minimum. We see that there is indeed an almost direct relation in the bias towards relevance in discovery and network communication. The right graphic provides a complementary view of this trend, where each line corresponds to a value of $p(tell|seen)$, the x axis is $p(tell|seen, relevant)$, and the y axis is the resulting $p(relevant|seen)$. Again, we see that the discovery/relevance bias grows with the sharing/relevance bias: the curves have monotonic growth with $p(tell|seen, relevant)$.

This correspondence between biases is not necessarily trivial, as we just noted earlier. The explanation is that, intuitively, in a relevance-prone sharing situation, items that many users like find more paths with high traversal probability (through connected friends who all like the item), and will therefore travel farther than items fewer users like, whereby more-liked items become discovered by more users.

4.2.2 Effect on popularity

We now check the effect of communication biases on popularity precision. We do so by the same parameter settings as we just did. Figure 3 shows the effects for the two variants of popularity-based recommendation described in section 3 (simple popularity on the top and relevant popularity at the bottom), in terms both the observed (left block) and true (right block) precision –more specifically, the difference between the precision of popularity and random recommendation. Similarly to Figure 2, for each popularity / precision variant we display a) a color map where the color scale shows the evolution of precision with respect to the user propensity to rate liked and non-liked items, and b) a graphic with a set of curves showing how precision evolves with the propensity to share liked items for a fixed amount of global communication $p(tell|seen)$ on each curve.

We see that popularity is in general better than random recommendation in most configurations, variants and metrics. However, we see that we may not take this for granted in all cases, and popularity becomes a worse than random approach in some circumstances. Comparing the graphics of true and observed precision, we see that the latter shows much lower values than the former. This is because observed precision only counts observed relevance in the form of ratings, which is a fraction of the total relevance that true precision takes into account –we are just reproducing the well-known fact that observed precision is a lower bound of true precision [7]. We may also observe that relevant popularity is generally a better option than simple popularity, as one would anticipate. We also see at first sight that the effects of the communication biases on the two popularity variants are almost the same in terms of observed precision, whereas they differ in terms of true precision, as we shall discuss next.

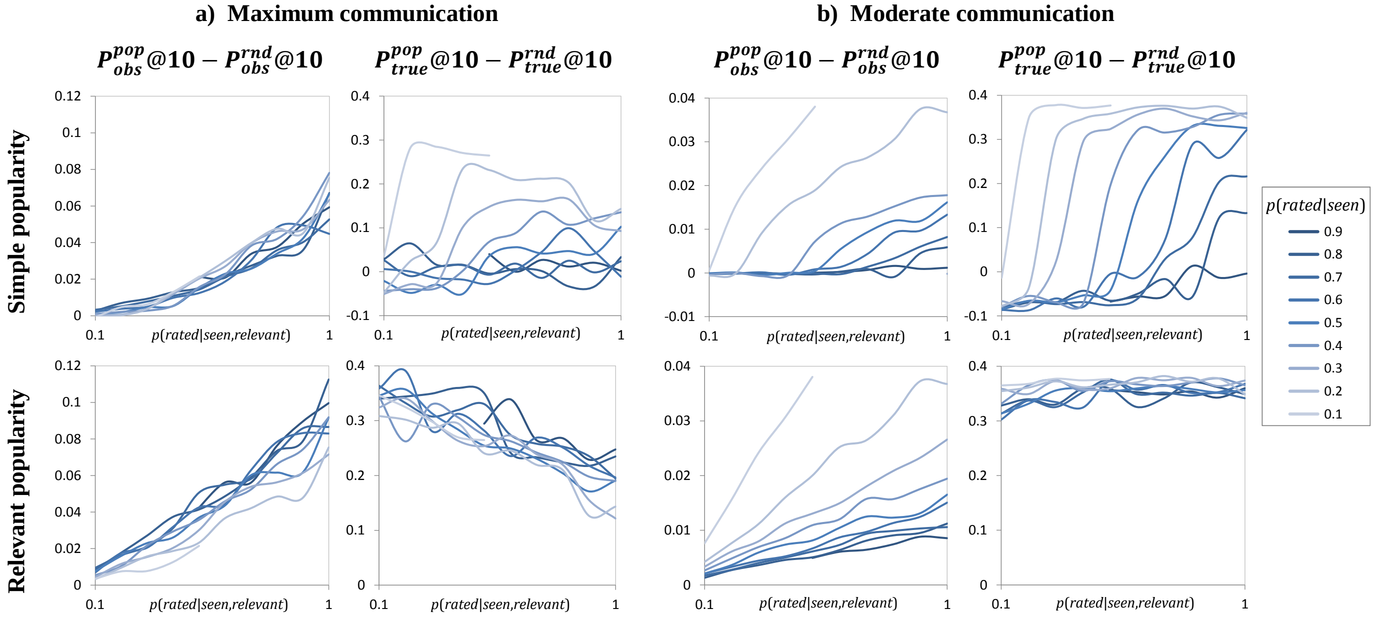

Figure 4. Effect of the rating/relevance bias on popularity precision. The relation between popularity precision and rating biases is shown for two different scenarios: a) intense communication (left) with $p(tell|seen, relevant) = p(tell|seen, \neg relevant) = 1$ and b) moderate communication (right) with $p(tell|seen, relevant) = p(tell|seen, \neg relevant) = 0.36$. Similarly to Figures 2 and 3, the curves at $p(rated|seen) = 0.9$ are truncated because it is not possible to have $p(rated|seen, relevant) < 0.5$ for such a high rating prior, and same for $p(rated|seen, relevant) > 0.5$ on $p(rated|seen) = 0.1$.

As a general trend, we see that popularity is more effective (in both variants and metrics) when users are prone to share items they like: all the rows of all the maps display a monotonic growth left to right, and all the curves in the graphics also show a steady growing trend. In terms of observed precision this is just natural: for a given number of total ratings (our simulation stopping condition), the number of positive ratings gets higher if discovery is biased towards liked items, and the difference to random precision is roughly proportional to the positive ratings density, for statistical reasons.

In true precision, the trend is explained because in a relevance-biased communication, the number of relevant and total ratings of each item will correlate with the number of users who like each item. Thereby liked items become statistically more popular, causing an increase in the resulting true popularity precision. In the opposite case, sharing items users do not like is counterproductive for popularity, because the generated ratings mislead the recommendation: the items with most ratings (predominantly negative) are not liked by many users, yet they get recommended by popularity. Simple popularity is particularly vulnerable to this, as it does not distinguish between positive and negative ratings. This popularity variant seems to be essentially sensitive to just whether $p(tell|seen, relevant) > p(tell|seen, \neg relevant)$ or the opposite. Once the inequality leans clearly enough to either side, precision does not vary much further, as we can see by the rather uniform colors in the triangular sections above and below the diagonal in the color map. In fact (though we do not show this information in the figure) the red and blue cells mostly correspond to negative and positive values respectively in this particular color map (i.e. precision is below or above random on each side of the diagonal).

Relevant popularity on the other hand is more robust: it is almost insensitive to the relevance bias for $p(tell|seen) < 0.5$ –the precision curves run almost constant and high with respect to $p(tell|seen, relevant)$ for such low levels of global communication. This makes sense because once popularity can correctly identify the relevant items by correlation with positive ratings, it does not matter how much rating information the recommendation is using for the prediction –the ratings do not count on the true precision computation, but just the full true relevance. Even in some regions where $p(tell|seen, relevant) < p(tell|seen, \neg relevant)$, true precision is good because negative ratings are ignored by this variant. However, sharing non-liked items beyond a certain degree does hurt relevant popularity: we see a strong decreasing trend in the color map columns. This happens because items which are not liked by that many users get enough positive ratings (corresponding to the few users who do like the items) to surpass highly liked items for which relevance remains more unobserved.

The trends on negative communication display some nuances in observed precision, where at some points sharing negative experiences seems to improve the measured precision of popularity. We hypothesize that this is due to the fact that when an item is shared, it may be liked by the receiver even if the sharer did not like it. Once negative discovery is saturated, further negative sharing may cause a slight relative raise in positive discovery (and therefore positive ratings), as we see in the color maps for $p(tell|seen, relevant) < 0.5$ in observed precision. We can also see that more communication results in better observed precision in the curves of the graphics (darker curves run above lighter ones). The trade-off between negative communication and total communication explains the "mix of diagonal trends" in the color maps in observed precision.

4.3 Rating/relevance bias

In order to study the effect of the rating bias on popularity recommendation, we take fixed neutral values for the communication biases $p(tell|seen, relevant) = p(tell|seen, \neg relevant) = 1$, that is, users share with their friends all the items they find. In this setting, we vary $p(rated|seen, relevant)$ and analyze the resulting curves for fixed values of $p(rated|seen)$.

Figure 4a shows the result. The first obvious trend is that observed popularity grows with the bias towards rating relevant items. This is because for a fixed total number of generated ratings (the simulation stopping condition), the number of positive ratings

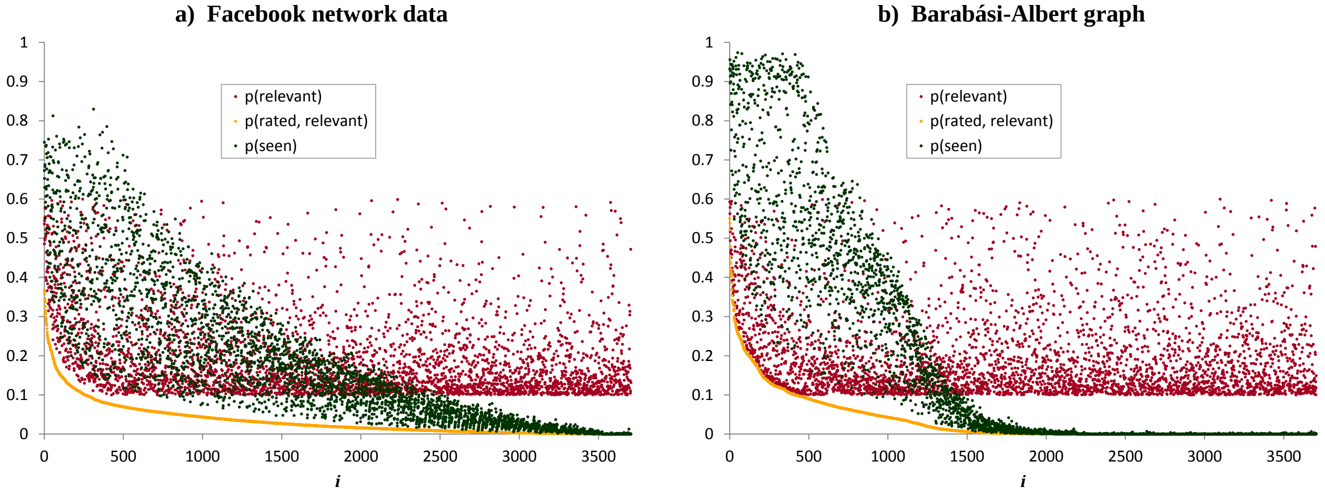

Figure 5. Positive rating, discovery and relevance distributions generated for different network structures: a) Facebook data (left) and b) a Barabási-Albert network with the same size in terms of number of nodes and arcs (right). The plots show the situation reached when the simulation has been run until 500,000 ratings are generated. The x axis corresponds to the items, sorted by their number of positive ratings. Each dot shows the ratio of users who like (red), have discovered (green), and have rated positively (yellow) the corresponding item. We may observe a steeper green discovery front (viral effect) in the Barabási-Albert network.

increases with that bias. And as pointed out earlier, this statistically increases the difference to random precision by roughly the positive rating density.

This effect is not observed in true precision. In fact, and quite paradoxically, the true precision of relevant popularity degrades with the positive rating bias (the curves display a decreasing trend). This is explained by a non-trivial interaction between a viral network effect and the recommendation protocol. The communication setting in these experiments is of extreme diffusion, since users always decide to share items every time they get a turn. Thus the first items to be found, by chance of exogenous discovery by some user, spread quickly through the network and accumulate a comparatively high number of ratings. The recommendation protocol mandates that users not be recommended items they have already rated themselves. This means that early rated items will be excluded from recommendations of more target users, as they got more ratings earlier. As $p(rated|seen, relevant)$ grows, generated ratings lean towards items with high $p(relevant|i)$. Thus the items excluded more often in the protocol will tend to be the most relevant ones, whereby true relevance decreases for statistical reasons: a decrease of the effective relevance density in the set of candidate items.

4.4 Network effects

As we have seen, in addition to the effect of individual users' behavior, further network effects may emerge from social-level dynamics, which end up affecting popularity. In order to complete the observation of viral effects just discussed in the previous section, we repeat the same experiments with lower communication rates $p(tell|seen, relevant) = p(tell|seen, \neg relevant) = 0.36$, which produces a positive rating distribution similar to MovieLens 1M –we omit the details about this for the sake of space.

Figure 4b shows the results. We see that the paradoxical effect of the positive rating bias in the true precision of relevant popularity disappears. Now in fact there is no dependence on the bias, as one should expect since relevant popularity ignores negative ratings, true precision ignores all ratings, and the correlation between relevance and positive ratings –which determines the effectiveness of relevant popularity– is achieved as soon as sufficient (a small number of) positive ratings have been generated. The exclusion of rated items is not that badly biased against relevant items because discovery being more evenly distributed, the generated ratings are not extremely skewed towards relevant items as before with viral propagation, and these "good items" are thus excluded from fewer recommendations.

The true precision of simple popularity does depend on the ratio of positive vs. negative ratings, and this is clearly shown in the corresponding graphic, where in fact precision steps up as soon as $p(rated|seen, relevant) > p(rated|seen, \neg relevant)$.

We also examine whether the network structure can be a factor in the observed dynamics. For this purpose, we run the same simulation on a Barabási-Albert (BA) graph [2] with the same number of users and friendship links as in the Facebook (FB) dataset. We take a simple configuration with $p(tell|seen, relevant) = p(tell|seen, \neg relevant) = 1$, $p(rated|seen, relevant) = 1$ and $p(rated|seen, \neg relevant) = 0$, that is, users share everything they discover, and rate only what they like.

Figure 5 shows the difference in the distribution of discovery and positive ratings on each graph. Discovery is steeper in BA than FB, with the best known items been discovered by almost all users, and the least known been discovered by almost none. Discovery is neutral with respect to relevance in this configuration (hence the green and red plots do not show correlation). The positive rating distribution correlates with discovery because the more an item is discovered the more chances it gets to be rated. It also correlates with relevance, because of the rating-relevance bias configuration.

Figure 6 shows the effect of this in the resulting precision. The results do not differ significantly between the two types of graphs, except mainly for the true precision of relevant popularity, which is good on FB, but close to random on BA. This is because information travels faster on the preferential attachment model of BA and the viral effect hurts true precision, whereas information takes longer to move outside friendship clusters on FB and popularity retains a better effectiveness. Note that in this setting, simple and relevant popularity are equivalent since all the generated ratings are positive as per the setting $p(rated|seen, \neg relevant) = 0$.

Popularity gets a very slight advantage in observed precision on BA. We attribute this to an effect of a steeper positive rating distribution with this graph, which results in the 10 topmost popular items accumulating a higher number of positive ratings, thereby yielding a higher observed $P_{obs}@10$.

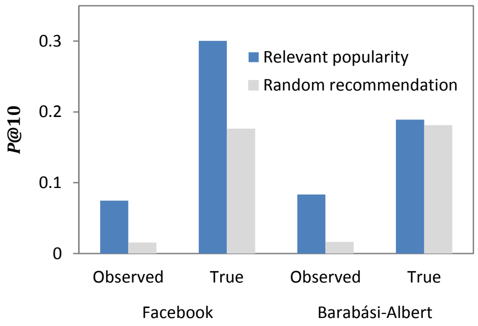

Figure 6. The effect of graph structure on popularity precision. The observed and true precision of relevant popularity vs. random recommendation on Facebook data (left) and a preferential attachment graph model (right) can be compared.

4.5 Effectiveness disagreement

The biases and effects we have analyzed can even cause a disagreement between observed and true relevance not just in quantitative terms, but also in terms of the comparison of two recommenders, in this case popularity and random. Figure 7 shows one such example on the FB graph. Simply with maximum “anti-relevance” communication bias $p(tell|seen, relevant) = 0$, $p(tell|seen, \neg relevant) = 1$, and neutral rating $p(rated|seen, relevant) = p(rated|seen, \neg relevant) = 1$, we get a negative direct correlation between positive ratings and relevance: the items with the most positive ratings are the ones with the least users liking them. As a consequence, the true precision of both simple and relevant popularity is worse than random. Oblivious of the full true relevance, the observed precision of popularity is still higher than random (which is too low to be barely visible in the figure).

5. CONCLUSION

We have studied the effect of different aspects of user behavior in social networks on the effectiveness of popularity-based recommendation. We define a formal model to represent such factors, based on which we can simulate different situations and observe the resulting effects. The analysis sheds findings such as a) popularity is most effective when users have a bias towards items they like when they share and rate items; b) highly active inter-user communication and viral information propagation pumps up observed precision –if communication is intense, even sharing non-liked items improves observed precision a bit further; c) viral propagation and a bias towards rating liked items can cause a decrease in true precision due to the exclusion of rated items from recommendations; d) viral propagation of negative opinions can cause a disagreement between measured and true precision even in terms of system comparisons; e) network structure is an additional factor for the effectiveness of precision, as it determines to a significant extent the speed of information transmission and discovery, thus intensifying or moderating the viral effects and their consequences on popularity.

The possibilities for continuation of the presented work are manifold. We are currently aiming to back the reported empirical observations with a formal analysis of the explored model and effects. Beyond that, the simplifications assumed so far can be relaxed in many directions: we may consider different inter-user communication modalities (e.g. communicate with more than one friend per turn, etc.), introduce the distinction between user discovery and item consumption, non-uniform user behaviors, biases in extrinsic item discovery (e.g. relevance bias in user searches), discovery loop from recommendations, temporal dynamics (e.g. new items and users keep entering the system), further dependencies between events (such as user decisions and choices depending on discovery source), dynamic networks, non-static user preferences (including effects of social influence in preference formation and propagation), etc. A user study to check what trends are given in practice in specific social environments would also be highly relevant to our research.

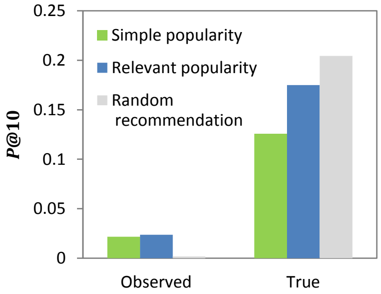

Figure 7. An example where observed and true precision disagree in the comparison of two recommenders (popularity variants vs. random recommendation).

6. ACKNOWLEDGMENTS

This work was supported by the national Spanish Government (grant nr. TIN2013-47090-C3-2).

7. REFERENCES

[1] Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. The Role of Social Networks in Information Diffusion. *WWW 2012*, Lyon, France, April 2012, 519-528.

[2] Barabási, A.-L. and Albert, R. Emergence of scaling in random networks. *Science* 286(5439), October 1999, 509–512.

[3] Blattner, M. and Medo, M. Recommendation Systems in the Scope of Opinion Formation: a Model. *Decisions Workshop at RecSys 2012*, Dublin, Ireland, September 2012, 32-39.

[4] Celma, Ò. and Herrera, P. A new approach to evaluating novel recommendations. *RecSys 2008*, Lausane, Switzerland, October 2008, 179-186.

[5] Cremonesi, P., Koren, Y., and Turrin, R. Performance of recommender algorithms on top-n recommendation tasks. *RecSys 2010*, Barcelona, Spain, September 2010, 39-46.

[6] Doerr, B., Fouz, M., and Friedrich, T. Why rumors spread so quickly in social networks. *Comm. ACM* 55(6), Jan 2012, 70-75.

[7] Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. Evaluating collaborative filtering recommender systems. *ACM Trans. Information Systems* 22(1), January 2004, 5-53.

[8] Marlin, B. M., Zemel, R. S., Roweis, S. T., and Slaney, M. Collaborative Filtering and the Missing at Random Assumption. *UAI 2007*, Vancouver, Canada, July 2007, 267-275.

[9] McAuley, J. J. and Leskovec, J. Learning to Discover Social Circles in Ego Networks. *NIPS 2012*, Lake Tahoe, NV, USA, December 2012, 548-556.

[10] Myers, S. A., Zhu, C., and Leskovec, J. Information Diffusion and External Influence in Networks. *KDD 2012*, Beijing, China, August 2012, 33-41.

[11] Pradel, B., Usunier, N., and Gallinari, P. Ranking with non-random missing ratings: influence of popularity and positivity on evaluation metrics. *RecSys 2012*, Dublin, September 2012, 147-154.

[12] Steck, H. Training and testing of recommender systems on data missing not at random. *KDD 2010*, Washington, DC, USA, July 2010, 713-722

[13] Steck, H. Item popularity and Recommendation Accuracy. *RecSys 2011*, Chicago, IL, USA, October 2011, 125-132.

[14] Steck, H. Evaluation of recommendations: rating-prediction and ranking. *RecSys 2013*, Hong Kong, October 2013, 213-220.