Identification of Influential Users in Emerging Online Social Networks Using Cross-Site Linking

Qingyuan Gong[1,2,3], Yang Chen[1,2,3], Xinlei He[1,2,3], Fei Li[1,2,3], Yu Xiao[4], Pan Hui[5,6], Xin Wang[1,2,3], and Xiaoming Fu[7]

[1] School of Computer Science, Fudan University, China
[2] Shanghai Key Lab of Intelligent Information Processing, Fudan University, China
[3] SKLCS, Institute of Software, Chinese Academy of Sciences, China
[4] Department of Communications and Networking, Aalto University, Finland
[5] Department of Computer Science, University of Helsinki, Finland
[6] CSE Department, Hong Kong University of Science and Technology, Hong Kong
[7] Institute of Computer Science, University of Göttingen, Germany
{gongqingyuan, chenyang, xinw}@fudan.edu.cn, yu.xiao@aalto.fi, panhui@cs.helsinki.fi, fu@cs.uni-goettingen.de

Abstract. Nowadays online social networks (OSNs) have become a commodity in our daily-life. Besides the dominant platforms such as Facebook and Twitter, a number of emerging OSNs have been launched in recent years, where users may generate less activity data than on dominant ones. Identifying influential users is critical for the advertisement and the initial development of the emerging OSNs. In this paper, we study the identification of potential influential users in these emerging OSNs. We build a supervised machine learning-based system by leveraging the widely adopted cross-site linking function, to overcome the limitations of referring to the user data of a single OSN. Based on the collected real data from Twitter (a dominant OSN) and Medium (an emerging OSN), we show that our system is able to achieve an F1-score of 0.701 and an AUC of 0.755 in identifying influential users on Medium using the Twitter data only.

Keywords: Influential Users, Emerging Online Social Networks, Cross-Site Linking

1 Introduction

Nowadays online social networks (OSNs) have become extremely popular around the world, and have attracted billions of users [13]. Besides supporting the interactions between people, OSNs also become platforms for information diffusion. The concept of *social influence* has been proposed to quantify the impact of different users within a selected OSN. Each OSN has a number of influential users, who could achieve a higher impact than most of the users. As discussed in [3], an influential user could be a "cool" teenager, an opinion leader or a popular public figure. Identifying influential users is very useful for different practical scenarios, such as viral marketing and making agile action to critical events. Several

social influence metrics have been proposed [3, 15] and evaluated on dominant OSNs such as Twitter. Meanwhile, many of the new (emerging) OSNs, such as Foursquare, Pinterest and Quora, are increasingly attracting registered users around the world.

An emerging OSN often offers some unique functions, instead of aiming to replace dominant OSNs such as Facebook and Twitter. For example, Foursquare provides location-centric services, Pinterest allows social content curation, and Quora acts as a question-and-answer site. Different from dominant OSNs, which offer general-purpose services, an emerging OSN typically has a special focus and only records each user's activities from limited aspects. Also, the account ages on emerging OSNs are in general younger than the account ages on dominant OSNs for same users. For example, according to our dataset, the average account age of Medium users is 3.39 years, while that of Twitter users is 7.70 years. As a result, it is harder to predict the social influence of a user on an emerging OSN due to the lack of comprehensive user activities. This problem is known as a challenging "cold start" problem [16], which appears when one user on a dominant OSN creates an account on an emerging OSN but has not added any information to her profile, or when she just plans to create an account on the emerging OSN. At this point, our goal is to predict whether she could become an influential user on the emerging OSN. Understanding the potential of the current less active users to be influentials is important to the advertisement and initial development of the emerging OSNs.

In this work, we study the problem of identifying influential users in "cold start" scenarios, i.e., predicting whether a newly registered user or a current inactive user on an emerging OSN would become an influential one. To solve the challenges brought by the inactiveness of the objective users, we leverage the power of the cross-site linking function [10]. We demonstrate that a user's rich demographic information and activity data on a dominant OSN can play an important role towards an accurate identification.

- We build an identification system to predict users' potential social influence in emerging OSNs. We take a story sharing social network, Medium, as an exemplified emerging OSN. We demonstrate how our system can make good use of user generated data on Twitter to achieve an accurate identification of influential users on Medium.
- We crawled the profiles and activities of 1 million Medium users, and their linked Twitter accounts. We conduct a data-driven study to evaluate our system. Our evaluation demonstrates that our system can achieve an F1-score of 0.701 and an AUC of 0.755, showing that cross-site linking can play an important role in identifying potential influential users on emerging OSNs.

2 Background and Data Collection

2.1 Cross-site Linking on Medium

Many of the emerging OSNs, such as Foursquare [6]/Swarm [5], Quora [18] and Pinterest [12], take advantage of their users' accounts on dominant OSNs to enhance their function-orientated services. They support a *cross-site linking function* [10], allowing users to link their accounts on dominant OSNs, e.g., Facebook and Twitter. In this way, users can log into the emerging OSNs with their Facebook or Twitter accounts, avoiding the trouble of managing multiple accounts. By enabling the cross-site linking function, a user can post the same piece of information to multiple OSNs simultaneously, copy social connections from dominant OSNs, and allow people to know more about her. By connecting a same user's accounts on multiple OSNs, the cross-site linking function provides opportunities to address the challenges in identifying influential users on emerging OSNs.

In this work, we select the "Medium-Twitter" pair as a case study, since Medium allows its users to link their profiles to Twitter. Medium is a blog sharing social network launched in August 2012. On Medium, each user maintains a profile page, showing her demographic attributes such as username and profile photo, her social connectivities including the number of followings and followers, and a paragraph of her biography. The linked Facebook and Twitter accounts are also shown on this page. Visitors to the Medium page can go to the users' Facebook and Twitter pages conveniently. A post on Medium is called a "story". If a user is impressed by a story published by other users, she can click a "Clap" button to show her appreciation or support. The profile page also provides links to three additional tabs, i.e., "Latest", "Claps", and "Responses", showing the latest stories published by the user, the stories she has clapped for, and the stories she has commented with, respectively. The stories in these tabs are organized with a reverse chronological order. In our investigation, Medium serves as the emerging OSN, and Twitter acts as the dominant OSN. Medium allows a user to link her Facebook and Twitter accounts to her Medium profile page. Users can only post tweets within 140 characters on Twitter, while Medium encourages users to post longer blogs without any character length limits.

2.2 Social Influence Definition

Users' influence on OSNs can be regarded as the power they can affect other users. The power can be quantified as P_u by various metrics depicting the users' activeness on the website and connectivity with others. For example, Kwak et al. [15] proposed three social influence metrics, i.e., number of followers, number of retweets and PageRank value. Based on the influence value P_u, the users can be categorized into two groups: the influential users and the ordinary users. Given a threshold p, the discrimination of these two user groups can be formularized as

$$u \text{ is } \begin{cases} \text{an influential user} & if\ P_u > p. \\ \text{an ordinary user} & if\ P_u \leq p. \end{cases} \tag{1}$$

The formularization is compatible with various metrics of social influence, since the formularization only takes the calculated value into consideration according to the definition. The threshold can be determined according to the specific requirement to the influential user. For example, it can be a percentage as top p% of the rank by the values of social influence, where a smaller p indicates that we select fewer users as influential users.

2.3 Data Collection

To obtain a dataset for our study, we need the activity records of Medium users and the data they generate on Twitter. We used Breadth First Search (BFS) to crawl the data of a number of Medium users, which has been widely used in OSN data collection, such as in [8,11,19]. We started our crawling from the user Evan Williams (https://medium.com/@ev), the CEO of Medium. In each step, we picked the first user from the head of the queue, recorded her user ID, and obtained a list of her followings and followers. These followings and followers, if have not been crawled, would be added to the queue. This procedure was repeated until the number of users we collected reached the threshold of 90 thousand. For each user in this dataset, we crawled her Medium profile and published stories. In our crawled data, 67.64% of the Medium users have linked their Twitter accounts. Accordingly, we further crawled these users' profiles and published tweets on Twitter.

3 Identification of Potential Influential Users on Emerging OSNs

The *social influence* of a user is an important concept in sociology and viral marketing [3]. In social networks, a piece of information can reach a large number of users through the network via a "word-of-mouth" way of diffusion. Since users may differ substantially in their credibility, expertise and social connectivity, they could achieve different levels of social influence. Some users, known as influential users, could quickly deliver information to a large number of audiences. Researchers have made numerous efforts in quantifying users' social influence in OSNs [1,3,15]. These studies focus on identifying influential users on Twitter, given the convenience brought by its global coverage and intensive user engagement.

In this section, we build a supervised machine learning-based system to discover the potential influential users on emerging OSNs in the cold start scenario, by leveraging the power of cross-site linking. We explain the overview of our model in Section 3.1, and present the implementation details in Section 3.2.

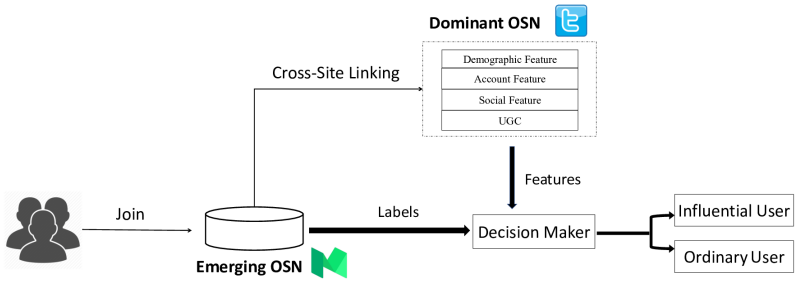

Fig. 1. Identification of Potential Influential Users

3.1 System Overview

We illustrate the system workflow by taking one user instance as an example in Fig. 1. Our goal is to predict whether this user is a potential influential user on Medium. For a user that enables the cross-site linking function and links dominant OSNs such as Twitter, we are able to access her Twitter account shown on her Medium profile page. By visiting the URL of her Twitter homepage, our system is able to use the collected information to determine whether this user will become an influential user on Medium. The data from the dominant OSNs allow us to have a better understanding of the user, extracted as a set of features to describe the user. A user's features will be fed into a decision maker, powered by a supervised machine learning-based classifier. Different supervised machine learning algorithms can be used to implement the decision maker. In this paper, we study classic algorithms include Random Forest [2] and C4.5 decision tree (J48) [17], as well as new algorithms including XGBoost [4] and LightGBM [14]. XGBoost and LightGBM are efficient tree boosting systems. Both of them have been widely used in machine learning contests such as Kaggle. The decision maker predicts whether the user will become an influential user, based on the extracted features of her.

3.2 System Implementation

Construction of the ground truth dataset A training dataset is needed to train a classifier to serve as the decision maker. For the user instances in the training dataset, there needs an explicit metric to determine the ground truth of whether each user is an influential user or an ordinary user, so that the classification performance of the decision maker can be evaluated. The ground-truth data is obtained based on the threshold defining the influential users. Our system is compatible with different kinds of metrics of social influence.

Extraction of feature sets Supervised machine learning-based classifiers need a number of features to describe an instance. Leveraging the cross-site link-

Table 1. Subsets of Features for Potential Influential User Identification

Site	Feature set	Feature List
Twitter	Demographic Features	· Length of biography · Has_added_location · UTC offset · Has_added_other_homepage
	Account Features	· Age of the account · Has_profile_image · Has_profile_background_image · Has_verified
	Social Features	· Number of followers · Number of followings
	UGC Features	· Number of tweets · Has_geo_tags · Number of lists subscribed to · Number of original tweets · Number of retweets · Number of "likes" sent · Number of "likes" received of original tweets · Average number of "likes" received of original tweets · Times of retweets of original tweets · Average times of retweets of original tweets

ing function, the system introduces illustrative features based on her publicly-accessible information on Twitter. We list our selected features in Table 1. Considering the services provided by OSNs, we classify users' features into 4 categories, i.e., demographic, account, social and UGC (user-generated content).

On Twitter, a user can choose to fill the information fields in her profile page, including a biography to describe herself, her current location, and the URL of her homepages. We also extract the information that describe a user's Twitter account, including the age of the account, the UTC offset she sets up, whether the user has changed the default profile image and the background image, and whether this account has been verified as an account of public interest. Since Twitter supports the social networking function by endowing a user to follow anyone she is interested in, we select the number of followers and followings as the social features. Besides receiving tweets from her followings or lists subscribed, Twitter users can also post tweets, publishing original tweets and re-tweeting other users' tweets. We extract comprehensive UGC features from users' tweeting behavior, including her preference of enabling the geography tags for tweets or not, her activeness of posting tweets, and the attentions she have received from other Twitter users in the forms of "like" and retweet.

The above features are all in decimal or binary values. These values construct a numerical vector for each user finally, depicting the user from demographic, account, social, and UGC aspects from her linked Twitter account.

Generation of the decision maker Fed with the feature vectors, the classifier employs a selected supervised machine learning-based classifier to learn the correlations between the extracted features and the social influence of users. We use a training dataset to get a set of "best" parameters of the decision maker. Af-

terwards, the trained decision maker is able to make the judgement whether one user will become an influential user on Medium, by referring to her information on Twitter.

There are several metrics to evaluate the classification performance. In our system, we use four representative metrics, i.e., precision, recall, F1-score and AUC (Area Under Curve). Precision measures the fraction of users classified as influential users who are really influential users. Recall is the fraction of users who are accurately identified as influential users. F1-score is the harmonic mean of these two metrics. AUC denotes the probability that this classifier will rank higher of a randomly influential user than a randomly chosen ordinary user. The performance for a selected classifier should be affected by the set of parameters in the classification algorithm, which can be attained through parameter tuning. We apply a grid search to sweep the parameters space of the given classifier, and choose the set of the parameters that could help the classifier achieve the highest F1-score.

4 Evaluation of the Influential User Identification system

In this section, we implement the proposed prediction system of influential users on Medium and evaluate its performance. The implementation and evaluations are based on the crawled data described in Section 2.3.

4.1 Evaluation Setup

We first construct a training dataset and build the ground truth of the influential users. In the evaluation, we consider the total number of "claps" received by a Medium user to quantify her influence. The intuition is that the "claps" can better reflect the social interactions between users than considering the following relationships only. As discussed in [19], the social interactions on Facebook are significantly skewed towards a small fraction of each user's social contacts. As a result, we rank the total number of "claps" received by Medium user, and take the top 10% as the threshold to divide the influential users and ordinary users in the dataset. After determining the ground truth dataset, we obtain the training dataset containing 4,624 randomly selected influential users and 4,624 randomly selected ordinary users.

To test the identification efficiency separately, another Medium dataset is introduced, called test dataset. In this dataset, there are 1,156 randomly selected influential users and 1,156 randomly selected ordinary users. Fed with the feature vectors of the users in the test dataset, the decision maker can give its judgement of whether any of them are influential users or ordinary users. In our system, we use different supervised machine learning algorithms to empower the decision maker, and train the parameters used through the approach of grid search.

4.2 System Performance

Using the metrics of precision, recall, F1-score and AUC, we show the performance of the identification system on the test dataset in Table 2. From this table, we see that when we feed the Twitter features into classification algorithms, the F1-score of XGBoost reaches 0.701, and the AUC value is as large as 0.755. For LightGBM, the F1-score and AUC values are a bit inferior to XGBoost, reaching 0.670 and 0.727 respectively. The F1-score or AUC will be both 0.5 if we apply a random guess to see whether one user will be an influential user, which can serve as the baseline of the prediction. This demonstrates the usefulness of our system, in particular, the usefulness of involving users' generated content on Twitter through cross-site linking.

Table 2. Performance of the Identification System of Influential Users

Algorithm	Parameters	Precision	Recall	F1-Score	AUC
Baseline	-	0.5	0.5	0.5	0.5
Decision Tree	criterion=entropy, min_samples_split=0.5, min_samples_leaf=10, min_weight_fraction_leaf=0.0	0.627	0.633	0.630	0.656
Random Forest	criterion=gini, max_depth=10, max_features=20, min_samples_leaf=1, min_weight_fraction_leaf=0.0	0.668	0.698	0.682	0.744
LightGBM	learning_rate=0.05, min_child_weight=1, max_depth=0, num_leaves=15, subsample=0.4, colsample_bytree=1.0, boosting_type=gbdt, objective=binary:logistic	0.658	0.684	0.670	0.727
XGBoost	learning_rate=0.1, min_child_weight=1, max_depth=12, gamma=0, subsample=0.9, colsample_bytree=0.6, booster=gbtree, objective=binary:logistic	0.691	0.713	0.701	0.755

4.3 Feature Importance

To evaluate the contributions of the features involved in the system, we feed the trained decision maker with one subset of features at a time. In such case, the classification algorithm judges the users in the test dataset only through the corresponding feature subset. We show the performance of the identification system with each subset of features listed in Table 1. The prediction performance results are shown in Table 3. It can be found that the social and UGC features of Twitter play important roles in improving the system performance. Demographic features are also helpful in the identification system. These results show the usefulness of Twitter features in the identification of influential users, which are based on cross-site linking.

We conduct the χ^2 analysis [20] to further measure the discriminative power of each feature of the identification system. Fig. 2 shows that the top 8 features cover all four categories. The content curation features rank the top 3, with the number of followers as the fourth one. The demographic and the account features also affect the prediction performance.

Table 3. Contribution of each Feature Sets in the Identification System

Feature subsets	Precision	Recall	F1-Score	AUC
Baseline	0.5	0.5	0.5	0.5
+Demographic features	0.556	0.540	0.548	0.607
+Account feature	0.622	0.213	0.318	0.561
+Social features	0.626	0.588	0.607	0.669
+UGC features	0.630	0.645	0.637	0.696

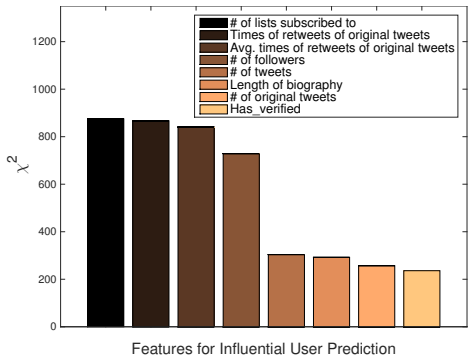

Fig. 2. χ^2 Analysis for the Identification System of Influential Users

5 Related Work

Identifying influential users is critical to study the information diffusion in OSNs. The Twitter platform has been widely used in studying the social influence. Cha et al. [3] explored three social influence metrics, i.e., in-degree (number of followers), number of retweets and number of mentions. These metrics represent different roles users play in OSNs. In-degree denotes the number of followers of a users, indicating how popular this user is concerning social connections. The number of retweets means how many times a user's tweets have been re-posted, showing the content value of her tweets. The number of mentions reflects the name value of a user. Existing explorations about social influence are mainly based on the rich activites of users on OSNs. For emerging OSNs, identifying the potential influential users is more challenging due to the lack of knowledge of users. In our work, we study the prediction of potential influential users on emerging OSNs like Medium. We leverage the cross-site linking function to introduce a user's information on established OSNs to realize the identification.

Cross-site linking is widely adopted by emerging OSNs as a way to take advantage of the established social connections of users on dominant OSNs. Zhong et al. [22] proposed the concept of social bootstrapping, i.e., copying existing friends from a dominant OSN into an emerging OSN. Their study demonstrated how a new OSN evolves by referring to the cross-site linking function. There are

research works trying to utilize the rich social footprints formed by users across OSNs. Zhang et al. [21] studied the relationship between social interactions on emerging OSNs and social ties on established ones. Farseev et al. [7] aggregated different kinds of information such as location, text, photo, and demographic attributes from Foursquare, Twitter, Instagram and Facebook to construct a multi-source dataset. They applied the dataset to predict users' demographic information. In our work, we make use of cross-site links between an emerging OSN and a dominant OSN. From this more informative view, we study how to discover potential influential users on the emerging OSN, even for the newly registered users.

6 Conclusion and Future Work

In this paper, we study the problem of identifying potential influential users on emerging OSNs. By introducing a user's usage and profile information on dominant OSNs, we propose a supervised machine learning-based system to predict whether she will be an influential user. Based on the real data of Medium and Twitter, we demonstrate an F1-score of 0.701 and an AUC of 0.755 of our system in distinguishing between influential users and ordinary users on Medium. Note that our system can be used by both OSN operators and third-party application providers, as the system only needs to access the publicly accessible information. For future work, we plan to expand our study to more emerging OSNs, and examine the prediction performance of our system. Moreover, we wish to study different types of user classification problems, such as the detction of malicious users [9], instead of merely focusing on identifying influential users.

Acknowledgement

This work is sponsored by National Natural Science Foundation of China (No. 61602122, No. 71731004), Natural Science Foundation of Shanghai (No. 16ZR1402200), Shanghai Pujiang Program (No. 16PJ1400700), EU FP7 IRSES MobileCloud project (No. 612212) and Lindemann Foundation (No. 12-2016), Projects 26211515 and 16214817 from the Research Grants Council of Hong Kong. Yang Chen is the corresponding author.

References

1. E. Bakshy, W. A. Mason, J. M. Hofman, and D. J. Watts. Everyone is an influencer: Quantifying influence on twitter. In *Proc. of ACM WSDM*, 2011.
2. L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
3. M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proc. of AAAI ICWSM*, 2010.
4. T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proc. of ACM KDD*, 2016.

5. Y. Chen, J. Hu, H. Zhao, Y. Xiao, and P. Hui. Measurement and Analysis of the Swarm Social Network With Tens of Millions of Nodes. *IEEE Access*, 6:4547–4559, 2018.
6. Y. Chen, Y. Yang, J. Hu, and C. Zhuang. Measurement and Analysis of Tips in Foursquare. In *Proc. of IEEE PerCom Workshops*, 2016.
7. A. Farseev, L. Nie, M. Akbari, and T. Chua. Harvesting multiple sources for user profile learning: a big data study. In *Proc. of ACM ICMR*, 2015.
8. N. Z. Gong, W. Xu, L. Huang, and et al. Evolution of Social-Attribute Networks: Measurements, Modeling, and Implications using Google+. In *Proc. of ACM IMC*, 2012.
9. Q. Gong, Y. Chen, X. He, Z. Zhuang, T. Wang, H. Huang, X. Wang, and X. Fu. DeepScan: Exploiting Deep Learning for Malicious Account Detection in Location-Based Social Networks. *IEEE Communications Magazine*, 2018.
10. Q. Gong, Y. Chen, J. Hu, and et al. Understanding cross-site linking in online social networks. *ACM Transactions on the Web*, 2018.
11. R. Gonzalez, R. Cuevas, R. Motamedi, R. Rejaie, and A. Cuevas. Google+ or Google-?: Dissecting the Evolution of the New OSN in Its First Year. In *Proc. of WWW*, 2013.
12. J. Han, D. Choi, B.-G. Chun, and et al. Collecting, Organizing, and Sharing Pins in Pinterest: Interest-driven or Social-driven? In *Proc. of ACM SIGMETRICS*, 2014.
13. L. Jin, Y. Chen, T. Wang, P. Hui, and A. V. Vasilakos. Understanding User Behavior in Online Social Networks: A Survey. *IEEE Communications Magazine*, 51(9):144–150, 2013.
14. G. Ke, Q. Meng, T. Finley, and et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proc. of NIPS*, 2017.
15. H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? In *Proc of WWW*, 2010.
16. P. Meo, E. Ferrara, F. Abel, and et al. Analyzing user behavior across social sharing environments. *ACM Trans. Intell. Syst. Technol.*, 5(1):14:1–14:31, 2014.
17. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.
18. G. Wang, K. Gill, M. Mohanlal, and et al. Wisdom in the Social Crowd: an Analysis of Quora. In *Proc. of WWW*, 2013.
19. C. Wilson, B. Boe, A. Sala, and et al. User Interactions in Social Networks and Their Implications. In *Proc. of ACM EuroSys*, 2009.
20. Y. Yang and J. O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *Proc. of ICML*, 1997.
21. P. Zhang, H. Zhu, T. Lu, H. Gu, W. Huang, and N. Gu. Understanding Relationship Overlapping on Social Network Sites: A Case Study of Weibo and Douban. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):120:1–120:18, 2017.
22. C. Zhong, M. Salehi, S. Shah, M. Cobzarenco, N. Sastry, and M. Cha. Social Bootstrapping: How Pinterest and Last.fm Social Communities Benefit by Borrowing Links from Facebook. In *Proc. of WWW*, 2014.