# Predicting the Link Strength of "Newborn" Links

Matteo Zignani
Dipartimento di Informatica
Universitá degli Studi di Milano, Italy
matteo.zignani@unimi.it

Sabrina Gaito
Dipartimento di Informatica
Universitá degli Studi di Milano, Italy
sabrina.gaito@unimi.it

Gian Paolo Rossi
Dipartimento di Informatica
Universitá degli Studi di Milano, Italy
gianpaolo.rossi@unimi.it

## ABSTRACT

Measurements of online social networks (OSNs) support the common fact that not all links carry the same social value, and that the strength of each link is strictly related to the frequency of interactions between the connected users. In this paper, we investigate the predictability of the interactions on OSN links by wondering if it is possible to categorize interactive or non-interactive links at their creation time. We turn the problem into a binary classification task and introduce a set of features which leverage the temporal and topological properties of the social and interaction networks, without requiring the knowledge of the interaction history of the link. The best classifier trained on a Facebook dataset obtained 0.72 as AUC. The above performance suggests that we can distinguish between interactive/non-interactive links at the time of link creation.

## Keywords

interaction prediction; interaction graphs; online social networks

## 1. INTRODUCTION

In the last years a great research effort has been spent to gather and measure users' relationships and their interactions through online social platforms including Facebook, Twitter, LinkedIn and Renren. These studies let emerge an across-the-board fact: not all links carry the same social value and the presence of a link between two users does not provide any information about their tie strength. Essentially the above social networks are an ensemble of links expressing strong and accidental friendships, acquaintances or even malicious relationships which need to be distinguished. To this aim a few studies [1, 2], while focusing on the estimate of the perceived tie strength through data from online social networks, have confirmed the fundamental role of the frequency of the interactions in determining a correct prediction of the link strength. Moreover the network among interacting users (interaction network) presents remarkable structural differences w.r.t the related social network in terms of small-worldness, hub nodes and outcomes of some graph-based applications [5].

Due to the above findings the interactions occurring on OSN links are becoming an important research topic, in particular their predictability has been partially investigated [3]. In this paper we ask whether the interactivity of two connected users can be predicted *a)* without requiring that users label links with a perceived tie strength; *b)* requiring that the prediction happens as soon as possible, i.e. at the creation of the link, namely imposing a zero-knowledge about the history of the interactions; and *c)* assuming no additional information except the timestamp for each interaction. Formally our formulation of the interaction prediction problem provides that, given a link $e_t(ij)$ created at time $t$ in the social network $G$ and a function $n^{\mathcal{T}} : E \to \mathbb{N}$ which returns the number of interactions on a link till the time $T$, we find a binary function $i$ such that:

$$i^{\mathcal{T}}(e_t) = \begin{cases} 0 & \text{if } n^{\mathcal{T}}(e_t) \leq \delta \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

The function $i$ must be applied at the creation time $t$ of the link $e_t$, i.e. the function will predict the interactivity of the pair $(i, j)$ as soon as they will connect.

The problem turns into a binary classification task, so we train and compare a set of binary classifiers that detect future interactive links at their creation. We have trained the classifier on a public Facebook dataset that includes temporal annotated relationships and interactions between the users. Due to the temporal constraint, the predictors cannot rely on the past interactions occurring on a link, unlike other proposed methods which exploit the interaction history [2]. This way we have to rely only on the topological (common neighbors, degree, clustering coefficient) and temporal features of the social network and a few properties taken from the interaction network at time $t$. The final results on the Facebook dataset show that the best classifier obtained 0.72 as AUC value and an accuracy equal to 0.65. The above outcomes suggest that we can distinguish between interactive/ non-interactive links at the link creation time.

Solutions for the interaction prediction problem would be useful to different social network applications. For instance, they could be apply to rank or filter user news feeds, or automatically setting the visibility of users' posts, or improve resource partition and allocation by estimating the volume of data access between users.


*WWW'16 Companion,* April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889367

## 2. DATASET

In this study we employ a network sample gathered from Facebook [4]. The dataset results from a crawling of the New Orleans Facebook network and captures the network growth of about 60.000 nodes and 800.000 links from September 2006 to January 2009. It contains the timestamped creation of users and edges, but 4.2% of vertices and 6.0% of links, which were not considered in our analysis. Moreover it contains user Wall interactions between 176054 pairs of users, corresponding to about 21% of the links in the network.

## 3. FEATURES

When coping with a supervised learning task, it is fundamental to identify a useful set of features on which to train the predictor. Our formulation of the interaction prediction problem relies on three main concepts: the social graph, the interaction network and the temporal information about their elements. This way we compute properties belonging to three different categories: topological features, interaction-graph features and temporal features.

**Topological features.** In this category fall all the measures which depend on the state of the social graph $G$ at the creation of a link. It involves the typical scores used in link prediction to assess the likelihood of a link to be created and metrics which estimate the centrality of nodes. Specifically, we favor computationally fast scores since they are computed every time a link is created. Given a link $e_t(ij)$, this category includes *a)* the $i$ and $j$'s clustering coefficient measured at the creation time $t$; *b)* the $i$ and $j$'s degree at time $t$; *c)* the Jaccard Coefficient $JC$ on $i$ and $j$'s neighborhoods at time $t$; and *d)* the Adamic-Adar Coefficient $AA$ on $i$ and $j$'s neighborhoods at time $t$.

**Interaction-graph features.** This category includes characteristics which depend on the interaction graph $I$. To this aim we define the neighborhood of a node $i$ in $I$ at time $t$, i.e. $\Gamma(i)^I_t = \{j | e_{t'}(ij) \in I, t' < t\}$. From the interaction graph we extract for each "newborn" link $e_t(ij)$ the following features:

1. $F1(i,j) = \frac{1}{|\Gamma(i)^I_t| + |\Gamma(j)^I_t|}$
2. $F2(u) = \frac{1}{|\Gamma(u)^I_t|}$, for $u = i, j$
3. Interaction propensity; $F3(u) = \frac{|\Gamma(u)^I_t|}{|\Gamma(u)^G_t|}$ for $u = i, j$
4. Interaction frequency; $F4(u) = \frac{n^t_\iota(u)}{t - t(u)}$ for $u = i, j$, where $n^t(u)$ denotes the number of interactions made by the node $u$ and $t(u)$ is $u$'s creation time.
5. Average interaction intensity; $F5(u) = \frac{n^t_\iota(u)}{|\Gamma(u)^I_t|}$, for $u = i, j$
6. Jaccard Coefficient; $JC(i,j)^I_t = \frac{\Gamma(i)^I_t \cap \Gamma(j)^I_t}{\Gamma(i)^I_t \cup \Gamma(j)^I_t}$
7. $F6(i,j) = \frac{\sum_{k \in \Gamma(i)^I_t \cap \Gamma(j)^I_t} n^t_\iota(ik) + n^t_\iota(jk)}{n^t_\iota(i) + n^t_\iota(j)}$

The features 1-2 relate to the centrality in $I$ of the link endpoints and their correlation, 3-5 capture the willingness to interact and the level of activity of the link endpoints, while 6-7 measure the embeddedness of the new link in $I$ considering or not the degree of interactivity of the endpoints.

**Temporal feature.** In [6] we introduce the *link delay* to indirectly measure the eagerness of a tie, obtained by quantifying the elapsed time between the potential establishment of a link and its real activation. The delay $d$ of the link $e_t(ij)$ is defined as $d(e_t) = t - \max(t(i), t(j))$. Lower the delay is, as soon as possible the two nodes will actualize the potential link.

| | Precision | Accuracy | AUC |
|---|---|---|---|
| Logistic | 0.654 | **0.707** | **0.723** |
| Random Forest | **0.678** | 0.672 | 0.722 |

**Table 1: Best classification results from the 10-fold cross-validation. Logit Boost, AdaBoost, Naive Bayes and C4.5 obtained worse results.**

## 4. RESULTS

We transform the interaction prediction into a binary classification problem. The negative class includes links with none or at most $\delta$ interactions from the creation of the link to the end of the sampling $T$, while the positive class contains the remaining links, the most interactive. To mitigate problems given by missing values we remove links where the delay is not defined and nodes which do not interact at all. Moreover we consider only links created an year before the end of the sampling.

In our case the complexity of the classification task is mainly given by the highly imbalanced classes. In accordance with the results in [5], the interaction network is more sparse than the related social network and consequently many links do not convey any interaction. To overcome this issue, we built a balanced training dataset by downsampling, thus keeping the positive instances and randomly picking an equal number of negative instances.

As partially shown in Table 1, we use the following classifiers: Logistic, Random Forest, Logit Boost, AdaBoost, Naive Bayes and a decision tree (C4.5). Since the dataset is imbalanced we evaluate the goodness of the classifiers in terms of accuracy, precision and AUC. In particular the last two metrics are suggested in literature in case of imbalanced classes. As reported in Table 1, the Logistic classifier obtains the best performance in terms of AUC and accuracy in the $\delta = 4$ setting. This way, given an active link, the classifier is able to correctly detect it with a probability greater than 0.65. In general the above results show that we can reasonably distinguish between interactive and non-interactive links at the time of the link creation.

## 5. REFERENCES


[1] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, New York, NY, USA, 2009. ACM.

[2] I. Kahanda and J. Neville. Using transactional information to predict link strength in online social networks. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM*, 2009.

[3] K. Kamath, A. Sharma, D. Wang, and Z. Yin. Realgraph: User interaction prediction at twitter. *User Engagement Optimization Workshop@ KDD*, 2014.

[4] B. Viswanath, A. Mislove, M. Cha, and P. K. Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM Workshop on Online Social Networks, WOSN 2009*, 2009.

[5] C. Wilson, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao. Beyond social graphs: User interactions in online social networks and their implications. *ACM Trans. Web*, 6(4):17:1–17:31, 2012.

[6] M. Zignani, S. Gaito, G. P. Rossi, X. Zhao, H. Zheng, and B. Y. Zhao. Link and triadic closure delay: Temporal metrics for social network dynamics. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.