Beyond Accuracy in Link Prediction

Javier Sanz-Cruzado and Pablo Castells

Universidad Autónoma de Madrid, Escuela Politécnica Superior
{javier.sanz-cruzado,pablo.castells}@uam.es

Abstract. Link prediction has mainly been addressed as an accuracy-targeting problem in the social networks field. We discuss different perspectives on the problem considering other dimensions and effects that the link prediction methods may have on the social network where they are applied. Specifically, we consider the structural effects the prediction can have if the predicted links are added to the network. We consider further utility dimensions beyond prediction accuracy, namely novelty and diversity. We adapt specific network, novelty and diversity metrics from social network analysis, recommender systems and information retrieval, and we empirically observe their effect on a set of link prediction algorithms over Twitter data.

1 Introduction

Link prediction can be considered today one of the classic areas in social network analysis research and development. The problem consists in finding links in a social network that have not been observed or formed yet, but may do so in the future, or may simply be useful to add. A paradigmatic application for the problem is recommending contacts in online social networks, a feature most social network applications, such as Facebook, Twitter or LinkedIn, nowadays provide.

A link prediction method can be evaluated in different ways, depending on the specific nuances in how the problem is stated. If seen as a classification task, the methods can be evaluated in terms of the predictive accuracy by usual metrics such as AUC, contingency tables, etc. [23]. If stated as a recommendation problem, information retrieval metrics can be used, such as precision, recall, etc. [15].

Yet as far as we are aware, most of the evaluation approaches to date, and therefore the solutions developed targeting them, seek to optimize a microscopic perspective of the network. In the classification perspective, a correctly classified link (true positive) adds as much to the accuracy metric as any other correct link, regardless where the people involved in the predicted edge are placed in the network, or what their social involvement may appear to be. Likewise, in the recommendation task, the metrics assess the accuracy or the benefit the recommendation brings to each target user in isolation, and then this microscopic benefit is simply aggregated into a "macro" average over all users.

Social networks are however not precisely about isolated users. It is well understood that a microscopic change (the formation of one link) has immediate direct and indirect effects in its surroundings; that small nearby changes combined produce something more than the sum of their effects; and that a few small changes spread across the network may have a substantial macro effect on the properties and behavior of the network as a whole. We therefore contend that we may want to consider the global consequence of link prediction on the network structure when assessing a link prediction algorithm.

We may moreover in fact want to target specific global effects in our link prediction methods. Contact recommendation functionalities nowadays account for an increasing fraction of the online social network growth; link prediction therefore represents an opportunity to favor trends towards desirable global properties in the evolution of a network, beyond (and in addition to) the short-term micro level value to be procured by the recommended links. The impact of recommendation can be quite important if we further consider the dynamic and recursive nature of social network growth.

The relation between link prediction and network evolution can be in fact obvious at the problem-statement level, and both views are naturally related in the literature: understanding how a network grows, and assessing the probability of any possible link to form, are very closely related tasks. Link prediction and network growth modeling have hence been seen as equivalent problems to some level. But how one (link prediction) can impact the other (network evolution) is yet to be studied; and the effect on network evolution has barely been considered, to the best of our knowledge, as part of the utility of the prediction algorithms to be evaluated.

In this paper, we discuss and explore several possible perspectives in the proposed direction. First, the social network analysis field provides a profuse array of concepts, metrics and analytic methods to assess the properties of the effect of link recommendation on a social network. We hence explore using such notions and measures to define new evaluation metrics for link prediction. Moreover, the recommender systems field has developed over recent years a clear awareness that accuracy alone is just a rather partial view on the value of recommendation: novelty and diversity perspectives can be as important –at both the macro and micro levels. We therefore likewise consider the adaptation of outcomes from that area to the evaluation of link recommendation. We find that, at more than one level, the global network analysis dimension of edge prediction links to similar principles as lay beneath the novelty and diversity perspectives.

2 Related work

Many different approaches to link prediction have been researched in the literature. Most proposed methods can be broadly classed in either of three categories: approaches based on the similarity between people [21], classical machine learning algorithms [22], and statistical network evolution models [13]. Link prediction can be applied to virtually any type of network, yet the problem has greatly gained importance with the explosion of world-wide online social networks, where the task is applied to recommend people to befriend or follow [12,14,15,31].

Link prediction and contact recommendation have so far essentially targeted the accuracy of the predictions. Incipient research has nonetheless considered the effects of contact recommendation algorithms on global properties of the network. We can distinguish two main perspectives in this scope. The first one focuses on the measurement of the effects of recommender systems on the structure of networks. The effect on metrics such as the clustering coefficient [8,17,30], the number of connected components [17] or the degree distribution [8] have been analyzed. The second line considers influencing the network growth towards some desired properties. In particular, Parotsidis [27] seeks to minimize the expected path length between the target user and the rest of the network; and Wu et al. [34] seek to maximize the modularity of the network. In this paper, we aim to broaden the perspective undertaken in such initial research, towards a wider range of network metrics, and dimensions beyond accuracy, such as novelty and diversity.

3 Notation

We shall use the following notation in the rest of the paper. We denote as $\mathcal{G} = \langle \mathcal{U}, E \rangle$ the graph structure of a social network, where $\mathcal{U}$ represents the set of people in the network, and $E \in \mathcal{U}_*^2 = \{(u,v) \in \mathcal{U}^2 | u \neq v\}$ represents the relations between them. We denote by $\Gamma(u)$ the set of people to which a person $u \in \mathcal{U}$ is connected. In directed networks, we shall differentiate between the incoming and outgoing neighborhoods $\Gamma_{\text{in}}(u)$ and $\Gamma_{\text{out}}(u)$.

The link prediction problem can be stated as identifying the subset of links $\hat{E} \subset (\mathcal{U}_*^2 - E)$ that are not observed but present in the network, or will form in the future, or would be useful to add –whatever the variant of the problem is. From a recommendation perspective, we shall denote by $\hat{\Gamma}_{\text{out}}(u)$ the set of people involved in the predicted arcs going out from u, i.e. $\hat{\Gamma}_{\text{out}}(u) = \{v \in \mathcal{U} | (u,v) \in \hat{E}\}$. And we shall refer to the graph including only the recommended links as $\hat{\mathcal{G}} = \langle \mathcal{U}, \hat{E} \rangle$.

4 Social network analysis

One way to assess the effect of a prediction algorithm on the network is to consider the extension of the network $\mathcal{G}' = \langle \mathcal{U}, E' \rangle$, with $E' = E \cup \hat{E}$, by a certain subset $\hat{E}$ of predicted links (for instance, the union of the top k predicted outgoing links in the ranking for each person u), as if the party the prediction is delivered to (e.g. the users of an online social network) accepted all the links in $\hat{E}$. Hence, any network metric applied to $\mathcal{G}'$ can be taken as a metric on the prediction method.

The social network analysis field is rich in metrics and concepts to characterize and measure network properties in many different angles. We summarize here some classical metrics we find of potential interest for the perspective under discussion. We suggest straightforward adaptations for our purpose, as well as further elaborations in some cases.

4.1 Distance-based metrics

An effect of recommendation, inasmuch as it increases the network density, is a general reduction of distances in the augmented graph. We may hence consider the metrics that measure this effect in different ways. Two common distance metrics are the average (ASL) and the maximum distance (diameter) over all pairs of people. We may also consider the farthest distance (eccentricity) for each person, averaged over all people. Moreover, in a recommendation perspective, we can measure the average distance between the people involved in predicted links in the original graph. We define reciprocal versions of the metrics when appropriate, in such a way that high values "are good" (in the sense that they reflect a possibly desired property).

- **Average reciprocal shortest path length:** $\text{ARSL}(\mathcal{G}') = \dfrac{1}{|\mathcal{U}|(|\mathcal{U}|-1)} \sum_{u,v \in \mathcal{U}} \dfrac{1}{\delta'(u,v)}$

  where $\delta'(u,v)$ denotes a shortest-path distance in the extended network $\mathcal{G}'$.
- **Reciprocal diameter:** $\text{RD}(\mathcal{G}') = 1 \Big/ \max_{u \in \mathcal{U}} ecc(u)$

  where the eccentricity $ecc(u) = \max_{v \in \mathcal{U}: \delta'(u,v) < \infty} \delta'(u,v)$ of a node u is defined as the distance to the farthest accessible node from u in the network [9].
- **Reciprocal average eccentricity:** $\text{RAE}(\mathcal{G}') = |\mathcal{U}| \Big/ \sum_{u \in \mathcal{U}} ecc(u)$

- **Mean prediction distance:** $\mathrm{MPD}(\hat{E}|\mathcal{G}) = |\hat{E}| \Big/ \sum_{(u,v)\in\hat{E}} \frac{1}{\delta(u,v)} - 2$

  where $\delta(u,v)$ denotes the shortest distance in the original network $\mathcal{G}$ (before prediction). MPD computes the harmonic mean of the prediction distances, subtracting 2 to set the metric minimum value at 0 (as $\delta(u,v) \leq 2\ \forall (u,v) \in \hat{E}$, since $\delta(u,v) < 2 \Rightarrow (u,v) \in E \vee u = v$, in which case the link is not considered for prediction: $(u,v) \notin \hat{E}$).

Distance shortening is likely a desirable effect in most cases, as it makes people easier to reach from each other through a smaller number of hops through common acquaintances. The different ways to average the distances provides nuances in the perspective with which distance is accounted for.

ARSL, RD and RAE range in $(0,1]$, and MPD takes values in $[0,\infty]$, reaching value ∞ when all the predicted links point to previously inaccessible people. It is easy to see that ARSL and MPD are defined in such a way that so-called global bridges [11], if any, between previously separate connected components are rewarded as the ideal case (where infinite distances are reduced to distance 1), whereas RD and RAE just ignore such improvements and just consider distance improvements within components. MPD measures in a quite direct way how the predicted links bring people far from their usual social environment, which can be seen as a measure of novelty from a contact recommendation perspective. We suggest the harmonic instead of the arithmetic mean because it can take in infinite distances.

4.2 Structural diversity

Notions of structural diversity have been a profuse object of study in the field of complex networks [20]. From the simplest perspective, the degree distribution can be seen as a primary sign of connective diversity: a very skewed distribution reflects a concentration of links around a few highly connected people, whereas in a flatter distribution each person has a more distinctive social circle of her own, and is exposed to different interactions than other people are. The "flatness" of the degree distribution can be summarized by a single number using the Gini index, which we can reverse into the **degree Gini complement**:

$$\mathrm{DGC}(\mathcal{G}') = 1 - \frac{1}{|\mathcal{U}|-1} \sum_{i=1}^{|\mathcal{U}|} (2i - |\mathcal{U}| - 1) \frac{|\Gamma'(u_i)|}{|E'|}$$

where people u_i in the above definition are ordered by non-decreasing degree $|\Gamma'(u_i)|$, and Γ' represents neighborhoods in the extended graph $\mathcal{G}'$. The same as we did for distance-based metrics, we take the complement of the Gini index, ranging in [0,1], in such a way high values indicate that the edges are evenly distributed, while values near zero indicate strong link concentration. In directed networks it makes also sense to compute indegree and outdegree versions IDGC and ODGC.

Richer notions of structural diversity have been studied, related to the concept of weak tie. Granovetter hypothesized that such links provide more novel information that strong ties [11]. In sociology, the strength of a link is defined as a combination of the amount of time spent on the relation, the emotional intensity, intimacy and reciprocal services that characterize the link. Strong links represent e.g. ties with family or close friends while weak edges correspond to more occasional acquaintances. Measures of link strength can also be defined based just on topological properties in the network, and can be related in some way or other to the sociological notion of weak tie [10,11]. Such measures are typically related to notions of redundancy: a tie is weak inasmuch as it is not redundant to other links around it; it carries a somehow exclusive (and hence valuable) connection between specific people or regions of the network. Such measures can be broadly divided in two categories: global and local.

Local notions. Local notions of weak tie consider the direct environment of a link to assess its strength. In this aim Granovetter considered a local weak link notion [13]: local bridges are links between people who do not have any common neighbors. We find this definition rather binary and restrictive, resulting in a coarse metric. The so-called link embeddedness provides a finer and more informative metric, which measures the relative overlap of the neighborhoods of its endpoints [36] as an indication of link strength:

$$\text{Embeddedness}(u, v|\mathcal{G}') = \frac{|\Gamma'_{\text{out}}(u) \cap \Gamma'_{\text{in}}(v)|}{|\Gamma'_{\text{out}}(u) \cup \Gamma'_{\text{in}}(v)|}$$

This metric actually smoothly generalizes the notion of local bridge: a link is a local bridge if it has embeddedness 0. We may assess the degree to which a link prediction method suggests weak ties by measuring the **average edge weakness** of the suggested links as the complement of embeddedness:

$$\text{AEW}(\hat{E}|\mathcal{G}') = \frac{1}{|\hat{E}|} \sum_{(u,v)\in\hat{E}} \left(1 - \text{Embeddedness}(u, v|\mathcal{G}')\right)$$

The metric ranges in [0,1] in such a way that the higher the weakness, the higher the structural diversity brought by link prediction.

Another classical means to assess the degree of connection redundancy around a person is the clustering coefficient, which can be defined as the ratio of neighbor pairs that are connected. The global clustering coefficient of a network can be measured by averaging this coefficient over all people in the network, or by an alternative global definition: the ratio of triangles in the network over the number of paths of length two. Again, we take the **clustering coefficient complement**, to get the metric values properly aligned with a notion of diversity:

$$\text{CCC}(\mathcal{G}') = 1 - \frac{|\{(u, v, w)|(u, v), (v, w), (u, w) \in E'\}|}{|\{(u, v, w)|(u, v), (v, w) \in E'\}|}$$

A link prediction method brings diversity to the network to the extent that this metric gets a low value.

Global notions. Along with the concept of local bridge, Granovetter proposed a global notion of bridge: a unique link between connected components. Again, we find this definition very restrictive in common social networks, which typically display a giant connected component [23], outside which the remaining components are rather marginal, and bridges between them are not particularly important to the network. We hence consider a relaxed definition based on work by De Meo et al. [10]: links between communities are considered weak ties, and links inside communities are considered strong. Inspired by this notion, we can consider different metrics that assess the presence of such links in the network.

A classical measure of the presence of inter-community links is the so-called modularity [24] (and reciprocally, many community detection algorithms consist in seeking partitions

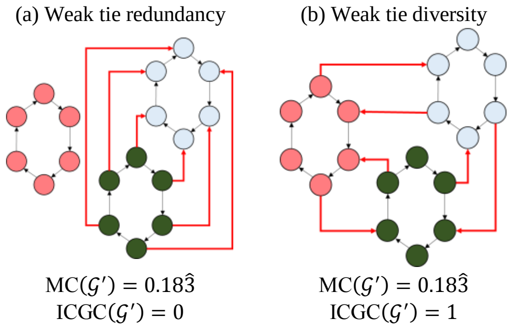

Figure 1. Networks with the same number of weak ties, but different link distribution. Black and red arrows represent, respectively, strong and weak ties. Nodes with the same color belong to the same community.

of $\mathcal{U}$ that minimize modularity [7]). Given a partition of the network into a set of communities $\mathcal{C}$, modularity compares the number of edges inside communities (strong links) to the expected number of strong ties we would find if the edges were placed at random:

$$\mathrm{Mod}(\mathcal{G}'|\mathcal{C}) = \frac{\sum_{u,v\in\mathcal{U}}(A_{uv} - |\Gamma'_{\mathrm{out}}(u)||\Gamma'_{\mathrm{in}}(v)|/|E'|)[c(u) = c(v)]}{|E'| - \sum_{u,v\in\mathcal{U}}|\Gamma'_{\mathrm{out}}(u)|\left|\Gamma'_{\mathrm{in}}(v)\right|/|E'|\,[c(u) = c(v)]}$$

where $c(u) \in \mathcal{C}$ denotes the community that u belongs to, A_{uv} is equal to 1 if there is a link between users u and v, and 0 otherwise, and $[\,\cdot\,]$ is the indicator function, which is equal to 1 iff the predicate inside the brackets is true. Once again, since low modularity indicates high diversity, we linearly reorient the values into a **modularity complement** metric ranging in [0,1], providing a direct measure of structural diversity:

$$\mathrm{MC}(\mathcal{G}'|\mathcal{C}) = \big(1 - \mathrm{Mod}(\mathcal{G}'|\mathcal{C})\big)/2$$

MC provides a measure of the abundance of weak ties across communities, but it does not provide information about the distribution of the weak links over the communities. The metric does not inform whether all those links connect a single pair of communities, or several different community pairs. The issue is illustrated in Fig. 1. We may hence want to consider a finer metric that assesses how balanced is the weak link distribution. For this purpose we propose the **inter-community Gini complement**, which counts the links between each pair of different communities, and computes the (complement of the) Gini coefficient of the distribution:

$$\mathrm{ICGC}(\mathcal{G}'|\mathcal{C}) = 1 - \frac{1}{M-1}\sum_{i=1}^{M}(2i - M - 1)p((c_1, c_2)_i|\mathcal{G}', \mathcal{C})$$

where $M = |\mathcal{C}|(|\mathcal{C}| - 1)$ is the number of pairs of (different) communities in the partition ($M = |\mathcal{C}|(|\mathcal{C}| - 1)/2$ if $\mathcal{G}'$ is undirected), $(c_1, c_2)_i$ is the i-th pair of communities with the smaller number of links between them, and $p\big((c_1, c_2)\big|\mathcal{G}', \mathcal{C}\big)$ is the probability of randomly selecting a weak tie between that pair of communities:

$$p\big((c_1, c_2)\big|\mathcal{G}', \mathcal{C}\big) = \frac{|\{(u, v) \in E'|c(u) = c_1 \wedge c(v) = c_2\}|}{|\{(u, v) \in E'|c(u) \neq c(v)\}|}$$

This metric has the extreme value 0 when only two communities have links across them, and 1 when every two pairs of communities have the same amount of crossing links. However, it does not inform of the total number of weak ties.

5 Novelty and diversity

Diversity is a rich concept that is studied in many different disciplines. Pertinent to our present focus, the information retrieval and recommender systems fields have developed notions of their own in this scope [2,4,6,32], which can be meaningful in the link prediction context as well. We hence consider their adaptation in a perspective where link prediction is seen as a contact recommendation task, targeted to the social network users.

5.1 Novelty

Novelty is a primary concern to recommender systems in most common scenarios where recommendation is tied to a purpose of discovery [4]. The most common novelty notion refers to recommending minority items in the long tail of the popularity distribution. In our case this means predicting links to people with little or moderate social involvement. A **long-tail novelty** metric can be formalized as the prior probability that a random person in the network was not acquainted to some of the recommended people to some other random user. This probability can be estimated by the proportion of people linking (in the original network $\mathcal{G}$) to recommended persons v, that is, $p(\neg\text{known}|v) = 1 - |\Gamma_{\text{in}}(v)|/|\mathcal{U}|$. The metric is hence defined as:

$$\text{LTN}(\hat{E}|\mathcal{G}) = \frac{1}{|\hat{E}|} \sum_{(u,v)\in\hat{E}} \left(1 - \frac{|\Gamma_{\text{in}}(v)|}{|\mathcal{U}|}\right)$$

which is inversely equivalent to the average indegree of the predicted contacts. This metric was proposed as the *expected popularity complement* in the context of recommender systems [5,32].

While LTN measures novelty from a global perspective (how novel are links to anyone), it also makes sense to consider the specific novelty from the individual viewpoint of each particular target user. So-called **unexpectedness** metrics have been proposed in the evaluation of recommender systems [4], which assess the dissimilarity between the recommended items and the prior experience of the specific target user. In our case the available records of user experience simply consist of their present contacts in $\mathcal{G}$:

$$\text{Unexp}(\hat{E}|\mathcal{G}) = \frac{1}{|\hat{E}|} \sum_{(u,v)\in\hat{E}} \frac{1}{|\Gamma_{\text{out}}(u)|} \sum_{w\in\Gamma_{\text{out}}(u)} d(v,w)$$

where the distance $d(v,w)$ between users can be defined in any meaningful way for the domain at hand. Measures based on path distance in the network would be somewhat redundant to the metrics suggested in section 4.1. Dissimilarity in terms of the social neighborhoods (e.g. Jaccard distance) is generally not the most informative option either, as it would typically represent a close opposite of the objective function of many prediction algorithms, and would hence tend to yield somewhat tautologically low values. Dissimilarity measures based on side-information about users are usually more meaningful.

LTN and unexpectedness can measure how much the recommended contacts takes target users far from their comfort zone, hence bringing opportunities for broadening and diversifying their social experience.

5.2 Diversity

Recommendation perspective. From a recommendation perspective, the diversity of a set of predicted links refers to how different are from each other the people recommended at the other end of the links, for a given target user. This is commonly measured by the **intra-list dissimilarity** [5], defined as the average pairwise distance between the people recommended to each target user:

$$\mathrm{ILD}(\hat{\mathcal{G}}) = \frac{1}{|\hat{E}|} \sum_{(u,v)\in\hat{E}} \sum_{w\in\hat{\Gamma}_{\mathrm{out}}(u)} \frac{d(v,w)}{|\hat{\Gamma}_{\mathrm{out}}(u)|}$$

where $d(v,w)$ is a dissimilarity measure between users, and can be defined in any meaningful way. Again, distance functions on user features tend to be more informative than measures based on the network structure around the users.

The second important diversity notion in recommendation concerns a global perspective and has often been referred to as the *diversity of sales* [5]. In our context it can be defined as how evenly are the recommendations distributed over all users, as opposed to being concentrated over a few people who are recommended to everyone. Again, the Gini index is a suitable summary metric to assess this aspect, which we use to define the **prediction Gini complement**:

$$\mathrm{PGC}(\hat{\mathcal{G}}) = 1 - \frac{1}{|\mathcal{U}|-1} \sum_{i=1}^{|\mathcal{U}|} (2i - |\mathcal{U}| - 1) \frac{|\hat{\Gamma}_{\mathrm{in}}(v_i)|}{|\hat{E}|}$$

where v_i represents the i-th user by non-decreasing number of times $|\hat{\Gamma}_{\mathrm{in}}(v_i)|$ she is recommended. This metric is equal to 1 when all users are recommended equally often, and 0 when all the predicted links point to the same single user.

Information retrieval perspective. A related but different take on diversity has been developed in search-oriented information retrieval, that considers returning diverse results considering the different possible intents or *aspects behind* an ambiguous search query [35]. Although link prediction does not involve explicit queries, it is possible to adapt this perspective by matching users in the network to queries and documents. The notion of aspect is abstract, and it can be particularized in many ways. We suggest considering latent communities as the equivalent of query aspects, whereby we can adapt all the aspect-based diversity metrics from IR.

The simplest metric, *subtopic recall*, which we may rename as **community recall** in our context, counts and averages the ratio of communities covered by the people recommended to each target user [37]:

$$\mathrm{CRecall}(\hat{\mathcal{G}}|\mathcal{C}) = \frac{1}{|\mathcal{U}||\mathcal{C}|} \sum_{u\in\mathcal{U}} \left| \bigcup_{v\in\hat{\Gamma}_{\mathrm{out}}(u)} c(v) \right|$$

where $\mathcal{C}$ is the set of communities and $c(v)$ is the community that user v belongs to.

A more elaborate approach to aspect-based evaluation is the so-called intent-aware scheme, in which one of the most meaningful and widely used metrics is **ERR-IA** [5]. This metric weighs down the added value of correctly predicted links in the ranking when the community of the recommended endpoint already occurs above in the ranking:

$$\text{ERR-IA}(\hat{E}|\mathcal{G}, E_{\text{test}}, \mathcal{C}) = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\sum_{c\in\mathcal{C}} p(c|u)\ \text{ERR-IA}(u,c)$$

$$\text{ERR-IA}(u,c) = \sum_{k=1}^{|\hat{\Gamma}_{\text{out}}(u)|} \frac{1}{k} p(\text{rel}|v_k, c) \prod_{j=1}^{k-1}\left(1 - p(\text{rel}|v_j, c)\right)$$

where v_k is the user at position k in the ranking of recommended links for user u, $p(\text{rel}|v,c)$ is commonly defined as $p(\text{rel}|v,c) = 0.5 \cdot [(u,v) \in E_{\text{test}} \wedge c(v) = c]$, and E_{test} represents the set of test links (held out from the prediction algorithms) with which the accuracy of link prediction is evaluated. The probability $p(c|u)$ that a community is pertinent to a user can be estimated by the ratio of followers of u that belong to c:

$$p(c|u) = \frac{|\{v \in c(u) | (u,v) \in E \cup E_{\text{test}}\}|}{\sum_{c'\in\mathcal{C}} |\{v \in c'(u) | (u,v) \in E \cup E_{\text{test}}\}|}$$

6 Empirical observation

In order to get a first feeling of the suggested metrics, we run a short experiment where we apply them to a set of state of the art link prediction algorithms over a sample of Twitter data.

6.1 Data

We use in our test the implicit dynamic network induced by the interactions between Twitter users, i.e. a link $(u,v) \in E$ if u retweets, mentions or replies v. We retrieve the data by a snowball sampling approach using the Twitter REST API. Starting from a seed user, we extract a selection of tweets published by that user. Then, we extract his social neighborhood as the set of outgoing interaction links present in the retrieved tweets. Then, all those users are added to the sample, and the procedure continues with the next retrieved user, until a desired number of users is retrieved. In our setup, we obtained 10,000 different users, and all their tweets published between the June 16th and July 16th 2015. The details of the collected dataset are shown in Table 1.

The link prediction task is set up for evaluation by means of a temporal split where the links dated prior to 9th July 2015 form the training set E of edges, and the links after that date are held out as test data E_{test}. The date of a link is defined as the timestamp of the first retrieved interaction between the users involved in the link. In fact, among all the metrics discussed here, only ERR-IA uses the test links –any pure accuracy metric, such as precision, naturally uses the test set as well.

We exclude reciprocating links from any prediction, as an obvious guess –the exceedingly high reciprocation ratio in Twitter would otherwise make the prediction task trivial; moreover Twitter already sends notifications anytime a new link is created (thus making an additional recommendation redundant).

For unexpectedness and ILD, we use as dissimilarity metric the complement of the *tf-idf* cosine similarity, using the concatenated text of the tweets posted by each user. Finally, for the community-based metrics, we tested different community detection algorithms, and we observed that the results are rather insensitive to the choice of algorithm (as already noted by De Meo et al. [10]). We will report the results using the Louvain method [3], one of the best known and most effective algorithms in the literature.

Table 1. Twitter dataset details.

# users $\lvert\mathcal{U}\rvert$	9,528
# training edges $\lvert E\rvert$	170,425
# test edges $\lvert E_{\text{test}}\rvert$	54,355

Table 2. Parameter settings for the algorithms

User CF	$k = 120$	MCN	$\Gamma_{\text{und}}(u), \Gamma_{\text{in}}(v)$
Item CF	$k = 300$	Pers. PR	$r = 0.4$
Implicit MF	$k = 260, \lambda = 150, \alpha = 40$	Matrix forest	$\alpha = 0.001$
SALSA	Authorities, $\alpha = 0.99$	Jaccard	$\Gamma_{\text{und}}(u), \Gamma_{\text{in}}(v)$
Local path	$\beta = 0.1, l = 3$	Katz	$\beta = 0.1$
Adamic	$\Gamma_{\text{und}}(u), \Gamma_{\text{in}}(v), \Gamma_{\text{und}}(w)$	Global LHN	$\lambda = 0.4$

6.2 Link Prediction Algorithms

We test the metrics on 15 different link prediction methods. Most of these algorithms have free parameters that we have tuned by simple grid search, targeting P@10; the resulting parameter settings are shown in Table 2 in the same notation as in [28]. We can classify the tested algorithms into four different families.

Neighborhood-based. We implement three algorithms in this group: one that ranks links according to the number of users in the intersection of neighborhoods [21] (most common neighbors, MCN), the Jaccard similarity of the users' neighborhoods [21], and the Adamic-Adar coefficient [1,21].

Path-based. We implement four different methods: the Katz algorithm [19,21] and a derivative algorithm, local path index [23]; the global Leicht-Holme-Newman (LHN) similarity index [20,23]; and the matrix forest index approach [27].

Random walks. We implement personalized PageRank [21,33], and the personalized SALSA algorithm [14] applied in the Twitter "Who-to-follow" service.

Recommendation algorithms. We adapt several recommendation methods to the link prediction task, including: user-based and item-based nearest-neighbor collaborative filtering (CF) [25]; a matrix factorization algorithm (Implicit MF) [16]; and a content-based approach that uses the text of tweets published by users [15].

In addition to all those algorithms, we include random link prediction, and popularity-based prediction, where the top most connected people are recommended to everyone.

As prediction thresholding method to select the set of predicted edges $\hat{E}$, we select the top k outgoing links as ranked by the predictor for each user in the network. In other words, we apply ranking cutoffs to the link prediction seen as a recommendation task targeted to the network users. To reflect this, we may append "$@k$" to all the metrics. In the results reported here, we take $k = \lvert\hat{\Gamma}_{\text{out}}(u)\rvert = 10$. Then, to compute the network analysis metrics, we add $\hat{E}$ to the original graph, as described earlier in section 4.

6.3 Results

We begin by observing in Figure 2 how the metrics correlate to each other, measured by the pairwise Pearson correlation over the set of 15 tested algorithms (the columns of Table 3). We easily spot several groups of metrics that seem to capture similar things.

ILD correlates with unexpectedness, which means that recommendations comprising a variety of users tend to be also different (on average) to the present contacts of the target user. AEW, CCC, and to lesser extent MC, are also correlated to each other, confirming the coherence of weak-tie and link-redundancy metrics. However, ICGC, also a weak-tie metric, seems to work in a different direction to the latter. We attribute this to

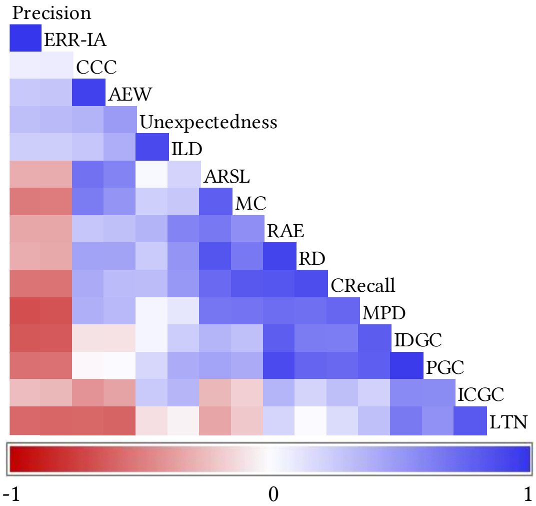

Figure 2. Pairwise Pearson correlation between metrics (@10 cutoff) for the values in Table 3.

the fact that long-tail recommendations (as measured by LTN) balance the distribution of cross-links over community pairs (as measured by ICGC), at the same time that LTN works against AEW and CCC as we discuss below.

Distance-based, Gini-based, community-based (including MC), and long-tail metrics also seem to roughly form a cluster of coherent measures, showing that they capture related sides of similar notions: connecting communities shortens distances between many users, as does avoiding link concentration. LTN and ICGC are however at the “periphery” of this cluster, and in fact correlate negatively with ARSL and MC. This is probably because connecting low-degree users (high LTN) does not shorten distances as much as linking to network hubs does. LTN seems to oppose CCC and AEW as well: links to low-degree users tend to be stronger than links to hubs, possibly because a) they contribute a smaller neighborhood size in the denominator of embeddedness for AEW, and b) long-tail people tend to have a low number of outgoing links as well, creating fewer unclosed triads for CCC.

Finally, precision correlates negatively with most novelty and diversity metrics, reflecting a general tradeoff already observed in the recommender systems field [4]. ERR-IA however takes only relevant recommendations into account, whereby it is known to strongly correlate with precision-oriented metrics. ILD, unexpectedness and AEW.

Table 3 shows the detailed metric values for all the algorithms –we also show in the top row, for reference, the metric values of the original (training) graph $\mathcal{G}$ when it makes sense (i.e. for metrics that do not use the set of predicted links $\hat{E}$). We start by noticing that the most accurate methods do not stand out, in general, in other perspectives. The most effective algorithm in precision, user-based collaborative filtering, scores very low for most novelty and diversity metrics (with the exception of ERR-IA, which correlates with precision). And the most inaccurate algorithm, random prediction, trivially achieves the most diverse network structure in terms of almost every metric.

The most outstanding approaches in long-tail novelty are the Jaccard coefficient, the Global LHN index and the content-based approach. Those methods, along with Matrix

Table 3. Metric values for a set of link prediction algorithms on the Twitter dataset described in Table 1. Column cells are colored from white (worst) to blue (best values), excepting the training graph, and saturating at the second highest value when random prediction has the top score (to avoid distorting the color scale). The best value of each metric is highlighted in bold. All metrics are at a @10 cutoff.

	Prec.	ARSL	RD	RAE	MPD	IDGC	AEW	CCC	MC	ICGC	LTN	Un-exp.	ILD	PGC	CRe-call	ERR-IA
Training graph	–	0.256	0.077	0.150	–	0.161	–	0.940	0.146	0.170	–	–	–	–	–	–
User-based CF	**0.068**	0.293	0.111	0.166	0.257	0.119	0.982	0.980	0.155	0.167	0.939	**0.931**	0.809	0.019	0.219	**0.087**
Implicit MF	0.063	0.293	0.111	0.148	0.378	0.117	0.980	0.977	0.155	0.178	0.957	0.871	0.766	0.029	0.238	0.077
Item-based CF	0.059	0.280	0.100	0.149	0.236	0.135	0.976	0.977	0.147	0.159	0.945	0.912	0.774	0.033	0.161	0.075
Pers. SALSA	0.058	0.316	0.125	0.159	0.387	0.112	0.985	0.982	0.166	0.167	0.928	0.861	0.745	0.011	0.238	0.072
Local path ind.	0.051	0.297	0.111	0.151	0.270	0.114	0.982	0.979	0.163	0.172	0.939	0.851	0.728	0.014	0.207	0.064
Adamic-Adar	0.051	0.301	0.125	0.168	0.211	0.135	0.969	0.975	0.149	0.170	0.958	0.846	0.762	0.056	0.223	0.060
MCN	0.048	0.299	0.125	0.155	0.207	0.132	0.971	0.975	0.146	0.169	0.959	0.833	0.737	0.045	0.204	0.057
Pers. PageRank	0.045	0.318	0.143	0.167	0.209	0.124	0.979	0.983	0.182	0.181	0.949	0.868	0.796	0.029	0.311	0.051
Matrix forest	0.039	0.288	0.111	0.161	0.163	0.166	0.955	0.974	0.160	0.182	0.979	0.869	0.764	0.120	0.266	0.046
Popularity	0.023	**0.357**	0.143	0.167	0.755	0.104	**0.997**	**0.989**	**0.295**	0.124	0.884	0.829	0.707	0.001	0.250	0.029
Jaccard	0.017	0.285	0.111	0.163	0.241	0.217	0.930	0.965	0.146	0.183	0.996	0.855	0.774	0.204	0.231	0.021
Content-based	0.013	0.316	0.111	0.147	0.731	0.253	0.948	0.965	0.156	0.190	0.997	0.890	0.710	0.233	0.238	0.018
Katz	0.010	0.296	0.111	0.151	0.706	0.116	0.985	0.984	0.261	0.180	0.974	0.900	0.784	0.014	0.331	0.013
Global LHN	0.001	0.299	0.100	0.141	0.978	0.221	0.958	0.977	0.146	0.170	**1.000**	0.784	0.628	0.085	0.179	0.001
Random	0.001	0.354	**0.200**	**0.241**	**1.929**	**0.439**	0.996	0.984	0.280	**0.200**	0.998	0.896	**0.838**	**0.823**	**0.497**	0.000

Forest, are also the only algorithms that improve the IDGC of the original network. The rest of methods seem to be biased towards linking to popular users, thus producing more skewed degree distributions in the extended graph. In terms of ILD and unexpectedness, nearest-neighbor collaborative filtering and Katz stand out. As one might expect, the content-based approach, which links similar users, achieves very poor values for these metrics.

We also observe that most methods (excepting random prediction) tend to connect close people in the network, as reflected in MPD. The algorithms that reach farthest nodes include popularity (highly connected people are not necessarily close to arbitrary target users) and path-based approaches such as Global LHN or Katz. Connecting to popular people also reduces the global distances in the network, as reflected by ARSL, RD and RAE –and PageRank has a similar effect, as it also tends to recommend hubs.

We can also see that links to hubs (popularity) tend to be weak according to AEW and CCC. This is because, on the one hand, AEW is heavily normalized by the neighborhood size, hence resulting in low values for popular people. On the other hand, indegree hubs tend to have a high outdegree; linking to them hence creates many new unclosed triads. This trend drags along the popularity-biased algorithms, such as the user-based CF, and certain path-based algorithms. However in terms of community-based metrics, collaborative filtering and neighborhood methods produce the lowest MC values, that is, the smallest number of community-based weak ties. This is to be expected, since near nodes are less likely to belong to different communities. The disagreement between MC and ICGC on popularity-based recommendation is most revealing as to the nuance of each metric. Links to hubs often cross communities and are hence seen as weak; yet the linked communities are most often the same few ones: the communities of the top k most connected (recommended) people, and hence the community connection diversity is poor according to ICGC. Both metrics otherwise often agree; for instance the Katz algorithm scores quite highly for both, and memory-based CF is rather low on both metrics.

7 Conclusions

Prediction accuracy seems like a rather partial perspective for link prediction considering the new (quantitative and qualitative) dimensions and role that social networks are acquiring, both as a service, a communication platform and a business. We find it natural to consider further perspectives when setting the target for link prediction and recommendation technology. In this paper we reflect on such perspectives and briefly explore some possibilities in this direction. Further metrics can be explored beyond the ones we test here. The metrics can be explicitly targeted either by devising new prediction algorithms that take the new dimensions into account, or by a post-prediction optimization of the output of an initial accuracy-oriented prediction algorithm, taking the desired metric as a second objective. This can be achieved by greedy re-ranking, multi-objective optimization, etc. [4,18].

While the benefit of accuracy is easy to motivate, we still need to understand better what the network perspectives discussed here imply in terms of their desirability and value for the people in the network, or any other concerned party (platform owner, network data consumers, etc.). For instance, shortening distances seems good for everyone: anyone can reach more people through fewer introductions by common friends [24]. Connecting distant people exposes them to the risk of an enriching experience. Enhancing the degree equality or promoting long-tail users helps avoid the disengagement of less involved people, and the saturation of hubs. Weak links may alleviate social bubbles [26] and/or enhance the speed and diversity of the information flow through the network [10,29,36]. Exclusive links between communities may bring strategic value [11], and so forth. This is certainly domain dependent, but can probably be studied also at some level of abstraction, which we envisage as future work.

References

1. Adamic, L.A., Adar, E.: Friends and neighbors on the web. Social Networks 25(3), pp. 211-230 (2003)
2. Agrawal, R., Gollapudi, S, Halverson, A., Ieong, S.: Diversifying Search Results. ACM WSDM 2009, pp. 5-14 (2009)
3. Blondel, V., Guillaume, J., Lambiotte, R., Lefebvrem E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics 10, P10008 (2008)
4. Castells, P., Hurley, N., Vargas, S.: Novelty and Diversity in Recommender Systems. In: Recommender Systems Handbook, 2nd Edition, pp. 881-918. Springer, New York (2015)
5. Chapelle, O., Shihao, J., Liao, C., Velipasaoglu, E., Lai, L., Wu, S.: Intent-based Diversification of Web Search Results: Metrics and Algorithms. Inf. Ret. 14(6), pp. 572-592 (2011)
6. Clarke, C., et al.: Novelty and diversity in Information Retrieval Evaluation. ACM SIGIR 2008, pp. 659-666 (2009)
7. Clauset, A., Newman, M., Moore, C.: Finding Community Structure in Very Large Networks. Physical Review E 70(6), pp. 1-6 (2004)
8. Daly, E.M., Geyer, W., Millen, D.R.: The Network Effects of Recommending Social Connections. ACM RecSys 2010, pp. 301-304 (2010)
9. Dankelmann, P., Goddard, W., Swart, C.: The Average Eccentricity of a Graph and its Subgraphs. Utilitas Mathematica 65 (2004)
10. De Meo, P., Ferrara, E., Fiumara, G., Provetti, A.: On Facebook, Most Ties are Weak. Communications of the ACM 57(11), pp. 78-84 (2014)

11. Granovetter, M.: The Strength of Weak Ties. American Journal of Sociology 78(6), pp. 1360-1380 (1973)
12. Goel, A., Gupta, P., Sirois, J., Wang, D., Sharma, A., Gurumurthy, S.: The Who-To-Follow system at Twitter: Strategy, algorithms and revenue impact. Interfaces 45(1), 98-107 (2015)
13. Guimerà, R., Sales-Pardo, M.: Missing and spurious interactions and the reconstruction of complex networks. PNAS 106(52), pp. 22073-22078 (2009)
14. Guy, I. Social Recommender Systems. Recommender Systems Handbook, 2nd Edition, pp. 511-543. Springer, New York (2015)
15. Hannon, J., Bennet, M., Smyth, B.: Recommending Twitter users to follow using content and collaborative filtering approaches. ACM RecSys 2010, pp. 199-206 (2010)
16. Hu, Y., Koren, Y., Volinsky, C.: Collaborative Filtering for Implicit Feedback Datasets. ICDM 2008, pp. 263-272 (2008)
17. Huang, X., Tiwari, M., Shah, S:. Structural Diversity in Social Recommender Systems. RSWeb Workshop at ACM RecSys 2013. CEUR Workshop Proceedings Vol. 1066 (2013)
18. Hurley, N., Zhang, M.: Novelty and Diversity in Top-N Recommendation – Analysis and Evaluation. ACM Transactions on Internet Technology 10(4), pp. 1-30 (2011)
19. Katz, L.: A new status index derived from sociometric analysis. Psychometrika 18(1), pp. 39-43 (1953)
20. Leicht, E., Holme, P., Newman, M.: Vertex similarity in networks. Phys. Rev. E 73(2) (2006)
21. Liben-Nowell, D., Kleinberg, J.: The Link-Prediction Problem for Social Networks. Journal of American Society for Information Science and Technology 58(7), pp. 1019-1031 (2007).
22. Lichtenwalter, R., Lussier, J., Chawla, N.: New Perspectives and Methods in Link Prediction. ACM KDD 2010, pp. 243-252 (2010)
23. Lü, L., Zhou, T.: Link Prediction in Social Networks: A Survey. Physica A 390(6), pp. 1150-1170 (2010)
24. Newman, M.: Networks: An Introduction. Oxford University Press, Oxford, UK (2010)
25. Ning, X., Desrosiers, C., Karypis, G.: A Comprehensive Survey of Neighborhood-Based Recommendation Methods. Recommender Systems Handbook, 2nd Edition, pp. 37-76. Springer, New York (2015)
26. Pariser, E.: The Filter Bubble. Penguin Books, New York, USA (2012)
27. Parotsidis, E., Pitoura, E., Traparas, P.: Centrality-Aware Link Recommendations. ACM WSDM 2016, pp. 503-512 (2016)
28. Sanz-Cruzado, J., Castells, P.: Contact Recommendations in Social Networks. In: Collaborative Recommendations, pp. 519-569. World Scientific Publishing, Singapore (2018)
29. Sanz-Cruzado, J., Castells, P.: Enhancing the Structural Diversity in Social Networks by Recommending Weak Ties. ACM RecSys 2018, pp. 233-241 (2018)
30. Su, J., Sharma, A., Goel, S.: The Effect of Recommendations on Network Structure. WWW 2016, pp. 503-512 (2016)
31. Tang, J., Hu, X, Liu, H.: Social Recommendation: A Review. Social Network Analysis and Mining 3(4), pp. 1113-1133 (2013)
32. Vargas, S., Castells, P.: Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. ACM RecSys 2011, pp. 109-116 (2011)
33. White, S., Smyth, P.: Algorithms for estimating relative importance in networks. ACM KDD 2003, pp. 266-275 (2003)
34. Wu, J., Zhang, G., Ren, Y.: A Balanced Modularity Maximization Link Prediction Model in Social Networks. Information Processing & Management 53(1), pp. 295-307 (2017)
35. Zhai, C, Cohen, W., Lafferty, J.: Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. ACM SIGIR 2003, pp. 10-17 (2003)
36. Zhao, J., Wu, J., Xu, K.: Weak ties: Subtle role of information diffusion in online social networks. Physical Review E 82(1) (2010).