Who will follow whom? Exploiting Semantics for Link Prediction in Attention-Information Networks

Matthew Rowe[1,2], Milan Stankovic[3,4], and Harith Alani[1]

[1] Knowledge Media Institute, The Open University, Milton Keynes, UK
[2] School of Computing and Communications, Lancaster University, Lancaster, UK
[3] Hypios Research, 187 rue du Temple, 75003 Paris, France
[4] Universit Paris-Sorbonne, 28 rue Serpente, 75006 Paris
m.c.rowe@open.ac.uk,mail@milstan.net,h.alani@open.ac.uk

Abstract. Existing approaches for link prediction, in the domain of network science, exploit a network's topology to predict future connections by assessing existing edges and connections, and inducing links given the presence of mutual nodes. Despite the rise in popularity of Attention-Information Networks (i.e. microblogging platforms) and the production of content within such platforms, no existing work has attempted to exploit the semantics of published content when predicting network links. In this paper we present an approach that fills this gap by a) predicting *follower* edges within a directed social network by exploiting concept graphs and thereby significantly outperforming a random baseline and models that rely solely on network topology information, and b) assessing the different behaviour that users exhibit when making followee-addition decisions. This latter contribution exposes latent factors within social networks and the existence of a clear need for topical affinity between users for a follow link to be created.

1 Introduction

Attention-Information Networks, or '*Hybrid Networks*' [10], lie at the intersection of social and information networks, users can *follow* other users and *subscribe* to the content they publish. Romero and Kleinberg [8] describe such directed interpolating networks as enabling users to become information hubs, in essence such users act as real-time sensors by disseminating information about real-world events and publishing information as it becomes available. Given the large uptake of platforms, such as Twitter (31.9 % increase in users in 2011[5]), that are composed of attention-information networks and the increased number of users to choose from, platform users must carefully select the individuals that they wish to *listen* to. Understanding *who will follow whom* and how users base their decisions - i.e. uncovering follower-decision behaviour patterns - has two key benefits: firstly, the dynamics of network growth in attention-information networks could

[5] http://www.emarketer.com/Article.aspx?R=1008879

be understood and therefore the social capital of the networks be predicted; and secondly, understanding how users behave in terms of their follower-decisions would facilitate audience building, a key interest for online marketing and brand managers who are keen to increase their broadcast spectrum.

Constrained attention capability means that users must decide on who they should follow. One would assume that a followee's content must be of interest to the follower, and this has indeed been identified in prior work by Schifanella et al. [9]. We therefore hypothesise that *Following a user is performed when there is a topical affinity between the follower and the followee*. However, to date no work has attempted to explore the differing behaviour that users may exhibit when making follower-decisions. We hypothesise that *Users who do not focus on specific topics do not base their follower decisions on topical information but on social factors*, as so-called **unfocussed** users who publish content about diverse subjects are not interested in subscribing to other users given a particular subject affinity. Further, we also hypothesise that *Users who are more socially connected are driven by social rather than topical factors*, given that users who build up a large followee network are more driven by connecting to people.

To explore these hypotheses we present an approach to predict links between a follower and recommended followees that exploits the semantics of user content, using tags and the concepts they refer to in order to measure the *semantic relatedness* of users. Our contributions are as follows:

- An approach to predict links in attention-information networks that explores *social*, *topical* and *visibility* factors, based on behavioural differences with regards to user types (alluded to in our hypotheses).
- Evaluation using the KDD Cup 2012 dataset from Tencent Weibo[6] that: a) shows significantly better performance than a random baseline and network topology models, and b) identifies a general pattern for follower-behaviour that is driven by topical affinity.

We have structured the paper as follows: Section 2 formulates our link prediction problem and describes recent work within this area. Section 3 describes the dataset used for our experiments. Section 4 details the prediction approach and the features engineered to capture social, topical and visibility dynamics, and Section 5 describes the method for concept disambiguation. Section 6 presents our experiments to identify follower-decision behaviour patterns and observe how users differ, and Section 7 discusses the findings in comparison with recent work. Section 8 finishes the paper with conclusions and plans for future work.

2 Background and Related Work

2.1 Problem Formulation

A social network can be modelled as a graph $G = \langle V, E \rangle$ where V denotes the set of users (nodes) in the social network and E is the set of edges ($\langle u, v \rangle \in E$)

[6] `http://t.qq.com/`

between nodes. Link prediction is the task of predicting which nodes $u, v \in V$ will form an edge between one another at a future time step. Leden-Nowell and Kleinberg [7] formulated this problem as detecting the changes in a graph between consecutive time steps ($t_0 < t'_0 < t_1 < t'_1$), by using information in $G[t_0, t'_0]$ to predict the edges in $G[t_1, t'_1]$. However, on attention-information networks the mechanisms through which edges, and therefore social links, are created requires that the link prediction problem is altered to account for recommendations - where a constrained set of possible nodes to connect to is considered. The introduction of recommendation features, such as the '*Who to follow*' feature on Twitter, has shifted the problem to a user-centric task such that a user u is provided with a set of recommendations $R(u)$ to connect to where $R(u) \cap \Gamma(u) = \emptyset$ and $\Gamma(u)$ denotes the ego-centric network of u. Therefore the problem we are addressing is the induction of a link function between users given previously provided recommendations: $f : V \times R \rightarrow \{0, 1\}$, where the set of possible mappings is constrained to the recommendation set of each user ($R(u)$).

2.2 Related Work

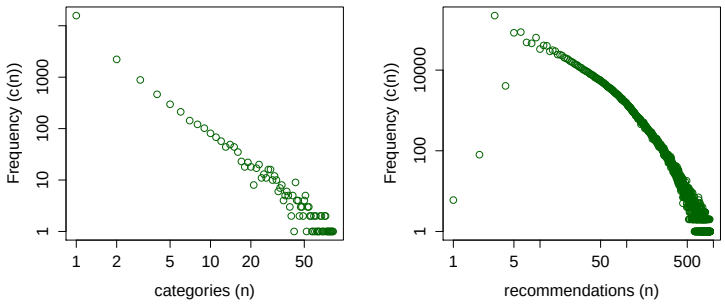
Recent work within the domain of link prediction is divisible into two strands: approaches that use network topologies and approaches that use local metadata. Starting with network topology driven methods, Golder et al. [5] modelled directed paths through networks to assess their effect on follower decisions on Twitter and found that increased transitivity (i.e. directed transitive connections) and common followers was correlated with follower addition. Also experimenting with Twitter, Yin et al. [10] found that 90% of created links are to users within 2 hops of a given user in the social network. Yin et al. assessed path structures through intermediate nodes and derived probabilities based on intermediary connections to predict links. Romero and and Kleinberg [8] examined '*directed closure*' (i.e. directed form of triadic closure) in attention-information networks and found that different link formation behaviour exists between sub-networks. Backstrom and Leskovec [1] proposed a supervised random walks method with restarts that combines node features with edge features to predict future links on Facebook. Edges are equivalent to the affinity between u and v in the context of our work and include features such as common friends. Zhou et al. [12] performed link prediction experiments over a range of network datasets ranging from a protein-protein interaction network through to a co-authorship network, where each network was undirected. The authors found common neighbours between nodes to achieve the best performance.

Focussing now on metadata-driven approaches, Schifanella et al. [9] assessed the correlation between tag affinity (overlap in tag vocabularies) and social neighbours for both Flickr[7] and Last.fm[8] users, finding that users who are close socially have common tags. Similar work by Leroy et al. [6] performed link prediction of Flickr users but with no *a priori* graph information, only using group

[7] `http://www.flickr.com`
[8] `http://www.last.fm`

membership information to indicate common interests. The authors' approach computed a probabilistic graph in a first *bootstrap* phase before using this information with existing topology-based measures from [7] to boost recall, finding that performance is favourable for common neighbours. Brzozowski and Romero [2] explored the effect of '*homophily*' on user recommendations, measuring this using the Dice coefficient of two users' sets of tags. However, contrary to findings in [9], Brzozowski and Romero [2] found that using similar tags was not useful information for predicting links and instead found mutual followers to be a good predictor. Yin et al. [11] predicted links using random walks with restarts by forming an augmented graph space of person nodes and attribute nodes, where attributes correspond to title keywords in an example DBLP co-authorship dataset, better performance was achieved when using local attributes (i.e. keyword information) rather than existing social connections.

Although existing approaches [9, 6, 2, 11] consider metadata when predicting edges between people the information is constrained to tag sets or group memberships and does not consider concept information. Furthermore, although several works indicate the benefit of using topical information [9, 6], there has been no examination of the follower-decision behaviour patterns. In this paper we present an approach that exploits semantics to gauge the topical affinity between a user and potential *followee* using concept graphs, and examine the decision patterns within our induced models.



(a) Categories per Tag (b) Recommendations per User

Fig. 1. Distributions in the dataset. Figure 1(a) shows that the distribution of categories per tag and Figure 1(b) shows the distribution of recommendations per user.

3 Dataset Description

The dataset that we used for the experiments described later in this paper was the KDD Cup 2012 dataset from the follower prediction task. Participants were given a rich collection of data that included: a) a training set of users with their recommendations, and whether they followed the items or not; b) the following graph of users; c) a set of keywords (tags) found within each user's content, and; d) item-categorisation data. This latter information is what we used for our concept hierarchy as a given item (i.e. a user) is placed within a hierarchical

concept graph - e.g. v is placed in 0.1.4.3 where the dot represents the branching of the concept hierarchy - and is also assigned several tags. This data has been manually labelled by the providers of the dataset. Both concept labels and keywords (tags) are anonymised and are replaced by numeric identifiers, so we do not see the underlying data. While the described concept hierarchy is used in our experiments, there is no obstacle to using some other concept hierarchy or structure (e.g. DBPedia) with our approach.

Figure 1(a) shows that many tags appear in a low number of categories, ($\mu = 2.275, \sigma = 5.903$), however certain tags (in the long tail) are extremely ambiguous and appear in many different categories, thus demonstrating the need for concept disambiguation. Figure 1(b) shows the distribution of recommendations per user ($\mu = 29.540, \sigma = 47.492$), with a skew towards lower recommendation counts and only a few users having a large number of recommendations.

4 Predicting Follower-Decisions

In this paper we tackle the problem of predicting links between a user and recommended *followee* users. We formulate the problem as one of inducing a function between the set of users and their recommendations: $f : V \times R \rightarrow \{0, 1\}$. Our goal is to explore various features and: a) identify the best performing general model over all users; and b) explore the decision behaviour of users and the extent to which this differs between them. In order to facilitate accurate predictions and explore the different factors that drive link creation we explore the use of three feature sets: *social*, *topical* and *visibility*. The features contained within these sets are each derived in a pairwise fashion such that if we are provided with a set of recommendations for u denoted by $R(u)$, we measure each feature based on the common information shared over u and $v \in R(u)$.[9]

4.1 Social

The decision of which recommendations to accept may differ between users, for instance it might be the case that one user may only *follow* another with whom they share a mutual friend. In order to assess such dynamics we measure four social features that account for the topology of the network and the existence of edges present within the network prior to predictions.

Mutual Followers Count Measures the overlap of the follower sets (i.e. the set of users connecting into a given user) between u and v. Let $\Gamma^-(u)$ denote the set of followers connected to u and $\Gamma^-(v)$ denote the set of users following v, then we define the mutual follower count as:

$$MFR(u, v) = |\Gamma^-(u) \cap \Gamma^-(v)| \tag{1}$$

[9] We use the symbols u and v hereafter to denote the user and a recommended user respectively.

Mutual Followees Count Measures the overlap of the followee sets (i.e. the set of users to whom a given user is connected) between u and v. Let $\Gamma^+(u)$ denote the set of followers connected to u and $\Gamma^+(v)$[10] denote the set of users following v, then we define the mutual follower count as:

$$MFE(u,v) = |\Gamma^+(u) \cap \Gamma^+(v)| \tag{2}$$

Mutual Friends Count Measures the overlap of the friends sets (i.e. the set of users with whom a user is friends, where friendship is denoted by a bi-directional edge between nodes) between u and v. The friend set is derived by taking the intersection of the followee and follower set of a given user. Using this set definition we can then calculate the mutual friends count as the overlap between friend sets between two users, or formally as:

$$MF(u,v) = |(\Gamma^-(u) \cap \Gamma^+(u)) \cap (\Gamma^-(v) \cap \Gamma^+(v))| \tag{3}$$

Mutual Neighbours Measures the overlap of the ego-centric networks of u and v whilst ignoring the directions of the links in the networks - this measure is taken from [12, 10, 1]. This feature is included to assess the impact, or lack of, that direction has on link creation - i.e. following or followed. We define this measure formally as:

$$MN(u,v) = |(\Gamma^-(u) \cup \Gamma^+(u)) \cap (\Gamma^-(v) \cup \Gamma^+(v))| \tag{4}$$

4.2 Topical

For certain users the decision to *follow* a user may be based on the content that the other user shares and produces. This effect is symptomatic of *attention-information networks* [5] in which the level of attention that a user can pay to content published by their network members is limited. To explore the effects of topical information on follower decisions we explore the overlap between users in terms of keywords (tags) and concepts. In the following section we describe a method to align a keyword to a concept given a user's context. Given such concepts we can explore the relation between users in terms of their semantic distance from one another within a concept graph, the intuition being that the further away two users are, then the less similar they are in terms of their interests, allowing the effect of *homophily* to be explored. We define several conventions as follows: let T_u be the set of keywords (or tags) found within the content of user u and C_{T_u} be the bag of concepts attributed to the tags from T_u.

Tag Vectors - Cosine Similarity Our first feature is similar to the cosine similarity between user tag vocabularies described in [9]. We define the tag vector $\boldsymbol{t_u} = \{t_1, t_2, ..., t_n\}$ of a user u as being derived from the user's tag set T_u using a binary index of the appearance of a tag within a user's content - i.e. $t_. = \{0, 1\}$

[10] We use the symbols $-$ and $+$ in the superscripts of the ego-centric networks to denote the direction of the edges, the former denoting incoming and the latter denoting outgoing.

in $\boldsymbol{t_u}$. To compute the similarity between the tag vectors of u and v denoted by $\boldsymbol{t_u}$ and $\boldsymbol{t_v}$ respectively we take the cosine of the angle between these vectors.

Concept Bags The concept bag C_{T_u} of a given user u is derived by returning the set of concepts that each tag in T_u has been associated with. As we have a collection of tags it is likely that duplicate concepts will be returned for different tags, we maintain these duplicates in the *concept bag* of a user and form a *concept bag vector*: $\boldsymbol{c_u} = \{c_1, c_2, ..., c_n\}$ using the frequency of the concepts in the bag as the weights - i.e. $c_{\cdot} = \{0\} \cup N^+$ in $\boldsymbol{c_u}$.

Cosine Similarity Given two concept bag vectors $\boldsymbol{c_u}$ and $\boldsymbol{c_v}$ for two different users u and v respectively, we measure the cosine of the angle between those vectors as the first measure between concept bags.

Jensen-Shannon Divergence Given two concept bag vectors $\boldsymbol{c_u}$ and $\boldsymbol{c_v}$ for two different users we model each vector as a probability distribution over the total set of concepts denoted as P_u and P_v respectively - using frequency counts for keyword usages to derive the probability distributions. Using these distributions over the total set of concepts we then measure the Jensen-Shannon Divergence between the concept bags, thereby gauging the level of dissimilarity between the concepts attributed to the content of u and v:

$$D_{JS} = \frac{1}{2} \sum_i P_u(i) \log \frac{P_u(i)}{P_v(i)} + \frac{1}{2} \sum_i P_v(i) \log \frac{P_v(i)}{P_u(i)} \tag{5}$$

Concept Graphs By using concept graphs we can explore the semantic relatedness of users using graph-based distance metrics. To enable this comparison we require a one-to-one mapping between a tag and a concept given a user. This produces a set of concepts for each user that can be used for comparison with other users. In the following section we explain how we perform *concept disambiguation* using a user's context to overcome the polysemy problem - i.e. where a single tag can have multiple concepts. We define $\langle t, c \rangle \in M_u$ as an injective map between the tags from the tag set T_u of user u and the set of concepts from the concept bag C_{T_u} of user u where a concept aligned to a tag given a user is returned by $M_u[t] = c$. Through this we can perform a pairwise comparison of the distances between concepts attributed to the tags of u and v using the function: $d(c_i, c_j)$. We define three distance measures over concept graphs - these distance measures are explained shortly - each of which have two varieties:

1. *Tag Intersection* The first variety uses the intersection of the tag sets of u and v for comparison: $T_u \cap T_v = T_{uv}$. For each tag in $t \in T_{uv}$ we produce the tag-concept maps given each user such that $|M_u| \equiv |M_v|$. The distances between the concepts from equivalent tags in T_{uv} are measured using a distance metric and the average taken. We define this formally as:

$$INT(T_{uv}) = \frac{1}{|T_{uv}|} \sum_i^{|T_{uv}|} d(M_u[t_i], M_v[t_i]) \tag{6}$$

2. *All Tags* The second variety performs a pairwise distance comparison between the tag sets of u and v through the concepts that each tag within those sets has been mapped to. For each tag in a given user's tag set T_u we produce an injective map: M_u. However, unlike the tag intersection the cardinality of the map for one user will differ from another if the cardinality of their tag sets differs. Using the maps we then measure the distance between every mapped concept in the different sets. We define this formally as:

$$ALL(T_u, T_v) = \frac{1}{|T_u|}\frac{1}{|T_v|}\sum_{i}^{|T_u|}\sum_{j}^{|T_v|} d(M_u[t_i], M_v[t_j]) \tag{7}$$

Based on these two varieties we explore three distance metrics for measuring $d(c_i, c_j)$ in the concept graph:

Shortest Path The first metric derives the shortest path between c_i and c_j using the Bellman-Ford algorithm. This method performs a breadth first search of a graph-space until a desired node is found.

Hitting Time The second and third metrics utilise the Markov-chain random walks model in which the probability of a random walker moving from one node to another in one time step is only dependent on their current position in a graph. The graph over which the random walker will traverse is the concept graph G_{conc} which is composed of nodes (concepts) V_{conc} and edges that connect those concepts $\langle i, j\rangle \in E_{conc}$ - where edges are undirected and therefore hypernym and hyponym relations are ignored for now. We define the random walks model using the Laplacian matrix of the concept graph: $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ and define the adjacency matrix $\boldsymbol{A}$ for entry a_{ij} to be 1 if $\langle i, j\rangle \in E_{conc}$ and 0 otherwise.

The diagonal degree matrix is defined as the row sum of the adjacency matrix: $d_{ii} = \sum_j a_{ij}$. We then take the Moore-Penrose pseudoinverse of the laplacian matrix which we denote as L^+. This provides, based on work by Fouss et al. [4], the necessary information to efficiently derive the hitting time $m(j|i)$ of a random walker as it traverses the concept graph G_{conc}, this is computed as follows:

$$m(j|i) = \sum_{k}^{|V_{conc}|} (l^+_{ij} - l^+_{ik} - l^+_{jk} + l^+_{kk})d_{ii} \tag{8}$$

Commute Time Distance The third distance metric computes the average number of steps the walker takes to leave a given node i reach another node j and then return back to i. The closer that two nodes are in the concept graph G_{conc} then the shorter the commute time. As the hitting time distances are not symmetric - i.e. $m(j|i) \neq m(i|j)$ - we define the commute time distance from i to j as: $n(j|i) = m(j|i) + m(i|j)$.

4.3 Visibility

The presence and access to information published by a prospective followee could influence users in deciding whether to follow the individual or not. However, limitations imposed on attention-information networks means that the dominant,

but not solitary, method through which posts by individuals outside of a user's followee network are seen is if they are *retweeted* or if a followee *mentions* a user. To explore these effects we devised the following six features:

Retweet Count The total number of times a given user (v) has been retweeted by members of the followee network belonging to u ($w \in \Gamma^+(u)$), we define this as follows, using the $retweet(w, v)$ function to return the number of times w retweeted v:

$$RC(u, v) = \sum_{w \in \Gamma^+(u)} retweet(w, v) \tag{9}$$

Mention Count The total number of times a given user (v) has been mentioned by members of the followee network belonging to u ($w \in \Gamma^+(u)$), we define this as follows, using the $mention(w, v)$ function to return the number of times w mentioned v:

$$MC(u, v) = \sum_{w \in \Gamma^+(u)} mention(w, v) \tag{10}$$

Comment Count The total number of times a given user (v) has had his/her content commented on by members of the followee network belonging to u ($w \in \Gamma^+(u)$), we define this as follows, using the $comment(w, v)$ function to return the number of times w commented on content published by v:

$$CC(u, v) = \sum_{w \in \Gamma^+(u)} comment(w, v) \tag{11}$$

Weighted Retweet Count The retweet count gauges the number of times a user v has been retweeted by members of the followee network $\Gamma^+(u)$ of u. The influence that members of this followee network exhibit may differ depending on the attention that u pays to each person. To assess this effect we set δ_w to be the number of times u has replied to $w \in \Gamma^+(u)$. We then derive a normalised influence weight λ_w for $w \in \Gamma^+(u)$ such that $\sum_{w \in \Gamma^+(u)} \lambda_w = 1$, where $\lambda_w = \delta_w / \sum_{w_\cdot \in \Gamma^+(u)} \delta_{w_\cdot}$. Given this influence weighting scheme we then measure the weighted retweet count such that the neighbours of u assert different effects on the count:

$$WRC(u, v) = \sum_{w \in \Gamma^+(u)} \lambda_w . retweet(w, v) \tag{12}$$

Weighted Mention Count As above, with the weighted retweet count, we also adjust the mention count of v by members of the followee network of u based on attention:

$$WMC(u, v) = \sum_{w \in \Gamma^+(u)} \lambda_w . mention(w, v) \tag{13}$$

Weighted Comment Count The comment count is also adjusted based on the attention paid by u to his followee network members:

$$WCC(u, v) = \sum_{w \in \Gamma^+(u)} \lambda_w . comment(w, v) \tag{14}$$

5 Concept Disambiguation with User Contexts

Measuring the distances between concepts within a graph space provides a notion of *semantic relatedness* that can, in turn, be used to cumulatively gauge the topical similarity between two users. As we mentioned above, distances are measured using three different metrics, however each metric requires an injective map between a set of tags and the concepts that they refer to. In this context we encounter the problem of concept ambiguity, also known as *polysemy*, where a single tag can have multiples concepts mapped to it.[11] Our earlier assessment of the distribution of categories per tags, as shown in Figure 1(a), demonstrates the large extent to which polysemy is evident within the dataset.

Cantador et al. [3] proposed '*distributional aggregation*' as a method for choosing the most representative tag for a web resource based on usage frequency amongst a collection of users. Our approach performs concept disambiguation by leveraging the context of the user, thereby swapping the collection of users for the concept bag of a given user and exploiting that as a *voting* mechanism. To illustrate this better consider a scenario in which the tag sets T_u and T_v and concept bags C_{T_u} and C_{T_v} are returned for for u and v. For each tag in the tag set we derive the list of candidates $C_{cand,t}$ for that tag (t) from the concept graph. For instance for a tag $t1$ we may have two candidates in the candidate set: $\{c1, c2\} \in C_{cand,t1}$. We count how many times each candidate appears in the concept bag of the user C_{T_u} and choose the most frequent, this then forms the mapping for the tag: e.g. $M_u[t1] = c1$. We define this process using the following function:

$$CD(C_{cand,t}, C_{T_u}) = \arg\max_c |\{c : c \in C_{cand,t}; c \in C_{T_u}\}| \tag{15}$$

6 Experiments

In the introduction of this paper we stated three hypotheses that describe the follower-decision behaviour of members of attention-information networks. The aforementioned features are engineered to capture the social, topical and visibility factors that could lead to follower decisions. In this section we describe experiments to verify our hypotheses tested over the KDD Cup 2012 dataset.

6.1 Experimental Setup

Our task, given that we are inducing a link function ($f : V \times R \rightarrow \{0, 1\}$), is a binary classification one. In essence we are asking *will user u follow user v?* To test our hypotheses we performed two experiments: *General Follower Prediction* and *Binned Follower Prediction*. For each experiment we first performed *model selection* by inducing a logistic regression model using only social, topical or visibility features and then all features combined together, we then selected the best model by the one that maximised the Area Under the ROC Curve (AUC).

[11] This is analogous to a word having multiple senses on Wordnet or the same tag appearing in multiple fixed taxonomical categories.

Second we assessed the coefficients in the logistic regression model trained on *all* features to identify patterns between an increase in a feature and the log-odds of the classifier increasing, and therefore the likelihood of a follower decision also increasing. The experiments were setup as follows:

General Follower Prediction Our first experiment sought a general follower prediction model in order to observe differences, at a general level, in follower behaviour. We randomly selected 10% of users from the dataset and generated a machine learning dataset for each user by building feature vectors $\boldsymbol{x_v}$ that contained the social, topical and visibility features (19 features in total) computed in a pairwise fashion between u and each recommended user $v \in R(u)$, setting the class label to *pos* if u followed v or *neg* otherwise. We combined the user-specific datasets together into one large dataset and balanced the data such that there were an equal number of positive and negative examples. The dataset, following balancing and a further randomisation process to ensure mixing, was then divided into an 80:20% split for training and testing, containing 457,722 instances in total.

Binned Follower Prediction We performed two experiments in this context. To begin with we measured two metrics for each user:

1. *Concept-bag Entropy:* We took the concept bag of each user derived from their tag set and measured the entropy of that concept bag, thereby capturing the dispersion of concepts that the user could be talking about. In this context **low entropy** denotes a **focussed** user while **high entropy** denotes an **unfocussed** user who is more random in the subjects that he publishes. We define this measure as follows, where $p(c_j)$ is the conditional probability of the concept (c_j) within the user's concept bag (C_{T_u}):

$$H_{C_{T_u}} = -\sum_{j=1}^{|C_{T_u}|} p(c_j) \log p(c_j) \tag{16}$$

2. *Degree Distribution:* We measure this as the proportion of users on the platform the user follows, thereby gauging how connected a user is. To derive this measure we took the out-degree of each user $(|\Gamma^+(v)|)$ and divided this by the total number of users $(|V|)$.

For each measure we divided the users up into 10 equal-frequency bins such that the same number of users were placed within each bin and selected all the users from the *low* and *high* bins. By choosing 10 bins and then selecting the low and high users for each of the above measures we are provided with users who will differ greatly in these properties. Following this binning process we built the datasets for each binned user using the same process as above (i.e. building pairwise feature vectors for each user u and each of his recommended users $v \in R(u)$) and then combined the user-specific datasets together, thereby producing four datasets for the experiment: two for the *Concept-bag Entropy* (*low* with 268,818 and *high* with 325,508 instances) and two for the *Degree Distribution* (*low* with 400,866 and *high* with 610,098 instances). We balanced each

dataset such that there were the same number of positive and negative instances and then divided each dataset up into a training/testing sets using an 80:20 split.

Evaluation Measures To assess the accuracy of the trained logistic regression models we measured the area under the Receiver Operator Characteristic Curve (AUC).[12] We use this measure to choose a model that predicts links accurately and minimises the number of false predictions. We also computed the Matthews correlation coefficient (MCC) to compare our models against a random predictor baseline: a coefficient of $+1$ is a perfect prediction, 0 is equal to random and -1 is total disagreement between prediction and observation.

6.2 Results: Prediction Accuracy

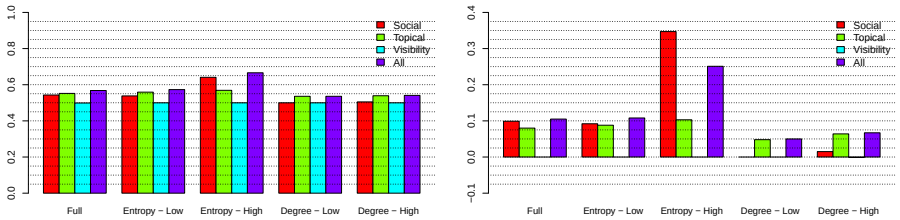
We begin our analysis of the prediction models by assessing the accuracy levels achieved in each experiment, as shown in Figure 2. Starting with the full model and assessing the feature sets used in isolation the results indicate that **topical** factors achieve the best performance and provide the most useful information for predicting links between users - performing significantly better (t-test with $\alpha < 0.001$) than social and visibility features. Within the introduction of this paper we hypothesised that '*Following a user is performed when there is a topical affinity between the follower and the followee*', the findings from our assessments of feature sets used in isolation confirms this hypothesis, however combining all the features together achieves the best performance. To provide an indication of the difference between the performance of the model and the baseline we ran the sign test of the MCC values and found each feature set combination to significantly outperform the random model.

Inspecting the AUC values produced for the Concept-bag Entropy models we find different patterns. For the *low entropy* users the **topical** factors perform best of the isolated feature sets, in a similar manner to the *full* model, while for the *high entropy* users the **social** features achieve the best performance of the isolated sets - significantly better than the other models (t-test with $\alpha < 0.001$). This finding confirms our earlier hypothesis that '*Users who do not focus on specific topics do not base their follower decisions on topical information but on social factors*'. Our intuition was that the more random a user is in his discussions then the less likely it would be for that user to base his follower-decisions on topical information. Instead, the driver in making such a decision is more likely to be social, as the user is more inclined to spread the topics and subjects he discusses in order to engage with more people.

Turning now to the Degree Distribution models we also find similar results to the full model by achieving the highest AUC values when using the **topical** features. Interestingly, we hypothesised earlier that '*Users who are more socially connected are driven by social rather than topical factors*', yet Figure 2 indicates that **topical** features outperform **social** features, thereby rejecting our earlier hypothesis - the former features were found to be significantly better (t-test with

[12] This accuracy measure is used throughout link prediction literature [12, 9, 6]

$\alpha < 0.01$). It could be the case that users who have a high out-degree form topic specific communities, assessing the coefficients of the logistic regression model for these high degree users should confirm this. We find that for all the models **visibility** features have little effect on predictions as only a small minority are non-zero - i.e. Retweet Count: $\mu = 13 \times 10^{-5}$.



(a) *AUC* (b) *MCC*

Fig. 2. Follower Prediction Model Results using the Full model and the Binned models. We report the Area under the ROC curve (*AUC*) and the Matthews Correlation Coefficient in parentheses, significantly outperforming the random baseline for all models bar **visibility** features.

6.3 Results: Follower-Decision Patterns

We now study the patterns of the logistic regression models in greater detail to compare the effects of various features on the log-odds ratio of the classifier and the probability of a user creating a link in their followee network. Starting again with the *full* model, and assessing the coefficients in Table 1,[13] we find that connections are formed between two users when there are fewer mutual followers, but more mutual neighbours - so they share some social affinity through their neighbours. Topically, users follow other users who are closer to them in terms of the subjects they discuss characterised by the reduced JS-divergence and greater cosine similarity across the tags and concepts, and also the reduced shortest path and hitting between *all* concepts of the users.

In terms of the binned models: *low entropy* users follow other users with whom they share less mutual followers but more mutual neighbours. These users should also have a greater topical affinity, given the negative coefficients for JS-divergence and the shortest path and hitting time across all concepts, and the positive coefficients for the tag vector and concept bag cosines. While *high entropy* users (who cover a lot of different topics in their discussions) follow other users with whom they share more mutual followers but less mutual friends. We also observe an interesting behaviour pattern for these user types as the tag vector cosine between a user and his followee should be minimised - indicating the requirement for a reduced overlap in the keywords that both users publish - however the concept vector cosine should be increased and the JS-divergence and hitting time should be reduced. As these latter features cover concept information, abstracted from tags published by either user, this suggests the presence of topical affinity between users without either user talking about the same tags.

[13] We only comment on features whose inclusion in the model is significant.

Table 1. Follower Prediction Model Coefficients for the General model (full) and the Binned models (Concept-bag Entropy and Degree Distribution).

		Full	Concept-bag Entropy		Degree Distribution	
Set	Feature		Low	High	Low	High
Social	Mutual Followers Count	-0.0275***	-0.0497***	0.2985***	8.6776	-0.0062**
	Mutual Followees Count	0.0001	-0.0064	0.1440***	-	0.0066***
	Mutual Friends Count	-0.0236	-0.1357	-0.2786***	-	0.0041
	Mutual Neighbours Count	0.0289***	0.0462***	-0.3033***	-	0.0023.
Topical	Tag Vectors - Cosine	0.7887***	0.5793**	-0.5125***	0.8628***	0.4840**
	Conc Bags - Cosine	0.6277***	0.9587***	1.6519***	0.5624***	0.5779***
	Conc Bags - JS-Divergence	-0.0410***	-0.0421**	-0.6369***	-0.0425***	-0.0059
	Conc Graphs - Int - Short Path	0.0329.	0.0811***	0.1324***	0.0556*	0.0795***
	Conc Graphs - All - Short Path	-0.0659***	-0.0444***	0.1515***	-0.0516***	-0.1230***
	Conc Graphs - Int - Hit Time	0.0009***	-0.0001	-0.0003**	-0.0002	-0.0002
	Conc Graphs - All - Hit Time	-0.0007***	-0.0006***	-0.0001.	-0.0005***	-0.0004***
	Conc Graphs - Int - Com Time	-0.0005***	0	0.0001	0.0001	0
	Conc Graphs - All - Com Time	0.0004***	0.0003***	0	0.0003***	0.0003***
Visibility	Retweet Count	4.3102	8.2279	6.9181	-	0.4570
	Mention Count	-12.5017	-	-	-	-2.4563
	Comment Count	-8.4571	-	-	-	-2.1373
	Weighted Retweet Count	-	-	-	-	-
	Weighted Mention Count	-	-	-	-	-
	Weighted Comment Count	-20.2386	-381.3810	-1401.6106	-	-1.1584

Signif. codes: p-value < 0.001 *** 0.01 ** 0.05 * 0.1 . 1

For *low degree* users we find that largely topical features appear within the model - as these users are not connected to many people and therefore the appearance of social factors is diminished. For these users the coefficients indicate a similar pattern for *low entropy* users where a user follows a recommended user if they share topical affinity - i.e. high cosine similarity based on tags and concepts, and lower hitting time and JS-divergence. *High degree* users follow other users with whom they share fewer mutual followers but more mutual followees - i.e. both users need to follow many of the same people. For topical features the high degree users follow other users with whom they share a topical affinity, indicating that although these users *subscribe* to many users they are based on common subjects and interest. We also find similar topical effects to both the *full* model and the *low entropy/degree* models: high similarity between the follower and followee based on tag vector/concept bag cosine, and low JS-divergence and hitting times.

7 Discussion

Analysing the follower-decision behaviour of users in an example microblogging platform proved two of our three earlier stated hypotheses. We also uncovered a general behaviour pattern based on the topical affinity between a follower and followee where the concept distance between the users should be lowered - measured using the random walks hitting time and tag similarity. Such findings are consistent with work by Schifanella et al. [9] where users who were socially close to one another were found to have a high topical affinity (cosine of tag vocabularies). Our work attempts to advance such findings by exploring novel metrics for assessing topical affinity, using concept graphs, and inspecting the coefficients

in induced logistic regression models to unearth the latent behavioural pattern. Such an examination has never been undertaken before.

In comparison with existing work we found different follower-decision behaviours. For instance, Golder et al. [5] found that on Twitter common followers, when increased in number, boosted the likelihood of a link being formed. This was also reported in Brzozowski and Romero [2] where sharing a mutual audience was correlated with better link prediction. However we do not see this effect in our general model and only see the increase in mutual followers as increasing the likelihood of a user following another in the *high entropy* model. In fact for the remaining models, the number of mutual followers should actually be reduced, thereby conflicting with both Golder et al. and Brzozowski and Romero's findings. Leroy et al. [6] found that an increase in mutual neighbours between a follower and followee was correlated with edge creation. We also observe a similar effect in our models where for all models, aside from the *high entropy* users, an increase in the number of mutual neighbours was associated with an increase in link creation likelihood. This divergent behaviour for *high entropy* users is common across many of the implemented features and suggests the need for model adaptation when considering these user types. Despite such divergent behaviour we note that a consistent topical affinity effect exists, as conceptually - i.e. considering concepts abstracted from published keywords - we find topical affinity between such random users and their followees.

8 Conclusions and Future Work

In this paper we have presented an approach to predict links on attention-information networks and in doing so: a) significantly out-performed a random model baseline when using all implemented features in a logistic regression model and existing topological models when using topical information; b) learnt a general pattern that captures the follower-behaviour of users of an example microblogging platform; and c) uncovered latent factors that lead to link creation including clear topical affinity between followers and followees. These findings allow followee recommendations to be improved based on the behaviour of the recipient, and therefore grow the network on the platform and increase social capital. A necessary next step for this work is to apply our models over data from other attention-information networks such as Twitter and YouTube in order examine the behaviour of their users and whether the findings from this work corroborate with those from disparate platforms. Assuming that concepts resolvable to Linked Data URIs can be extracted from textual content available in those networks, this would allow our concept-based affinity measures to be applied over Linked Data.

Future work will also involve assessing the correlation between social network distances between users and their topical affinity. Schifanella et al. [9] found when one compares users who are more than 2 steps away in a social network then the topical affinity between users declines rapidly. We plan to combine this examination with applying our approach over Twitter and YouTube. We are

also exploring the use of Conditional Random Fields on link prediction, thereby allowing follower-decisions to be conditioned on recent user behaviour - i.e. a user's recent propensity to follow other users. We conjecture that performance is conditioned on time-sensitive behaviour of each user. We do not capture this at present.

Acknowledgments

This work was supported by EU-FP7 project ROBUST (grant no. 257859).

References

1. Lars Backstrom and Jure Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proc. of the international conference on Web search and data mining*, WSDM '11, New York, NY, USA, 2011. ACM.
2. Michael J Brzozowski and Daniel M Romero. Who Should I Follow? Recommending People in Directed Social Networks. In *International AAAI Conference on Weblogs and Social Media*, 2011.
3. Iván Cantador, Alejandro Bellogín, Ignacio Fernández-Tobías, and Sergio López Hernández. Semantic contextualisation of social tag-based profiles and item recommendations. In *EC-Web*, pages 101–113, 2011.
4. Francois Fouss, Alain Pirotte, Jean-Michel Renders, and Marco Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Trans. on Knowl. and Data Eng.*, March 2007.
5. Scott A. Golder and Sarita Yardi. Structural predictors of tie formation in twitter: Transitivity and mutuality. In *Proc of the IEEE International Conference on Social Computing*, Washington, DC, USA, 2010.
6. Vincent Leroy, B. Barla Cambazoglu, and Francesco Bonchi. Cold start link prediction. In *Proc of the international conference on Knowledge discovery and data mining*, New York, NY, USA, 2010. ACM.
7. David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, 58(7):1019–1031, May 2007.
8. Daniel Mauricio Romero and Jon M. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *ICWSM*, 2010.
9. Rossano Schifanella, Alain Barrat, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Folks in folksonomies: social link prediction from shared metadata. In *Proce of the international conference on Web search and data mining*, New York, NY, USA, 2010.
10. Dawei Yin, Liangjie Hong, and Brian D. Davison. Structural link analysis and prediction in microblogs. In *Proc of the ACM international conference on Information and knowledge management*, New York, NY, USA, 2011.
11. Zhijun Yin, Manish Gupta, Tim Weninger, and Jiawei Han. Linkrec: A unified framework for link recommendation with user attributes and graph structure. In *Proc of the World Wide Web Conference*, New York, NY, USA, 2010.
12. Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. Predicting missing links via local information. *The European Physical Journal*, October 2009.