Predicting Interests of People on Online Social Networks

Apoorv Agarwal
Department of Computer Science
Columbia University
NYC, NY 10027
Email: apoorv@cs.columbia.edu

Owen Rambow
Center for Computational Learning Systems
Columbia University
NYC, NY 10027
Email: rambow@ccls.columbia.edu

Nandini Bhardwaj
Department of Computer Science
Columbia University
NYC, NY 10027
Email: nb2338@columbia.edu

***Abstract*—We introduce a new data set which contains both a self-declared friendship network and self-chosen attributes from a finite list defined by the social networking site. We propose Gaussian Field Harmonic Functions (GFHF), a state-of-the-art graph transduction algorithm, as a novel way of testing the relevance of the friendship network for predicting individual attributes. We show that the underlying self-declared friendship network allows us to predict some but not all attributes. We use Support Vector Machines (SVM) in conjunction with GFHF to show that other attributes such as age or languages spoken are also important.**

I. Introduction

Homophily [1] is a well-documented phenomenon: people choose friends who are, with respect to some characteristics, similar to them.[1] If homophily is a robust aspect of human behavior, then we can use it to deduce people's characteristics from their friends' characteristics, as long as the characteristics in question are among those that people take into account when choosing friends. (For example, political opinion may factor into the choice of friends, but handedness may not.) The growth of online social networking sites with explicit, binary, and mutually self-defined friendship networks provide us with new opportunities to explore the issue of homophily in an empirical manner. It also raises new questions related to homophily. In this paper, we focus on the following three questions.

- Are mutually self-declared friendship links on social networking sites compatible with the homophily assumption? This question arises as the self-declaration of friendship on a social networking site is a rather different cognitive activity from developing and maintaining an actual friendship, be it online or offline.
- Which personal attributes correlate with mutually self-declared friendship?
- Which statistical methods allow us to detect homophily and to exploit it for prediction?

There has been much work on inference using network information, some of which is relevant to our questions (see a review in Section II). We would like to immediately point out an important difference between our work and some of the other work about predictions in networks. In some related work, attributes are predicted from a network, but this network is not a mutually self-declared friendship network. For example, some networks are derived from the attributes of nodes. The network does not exist independently of the node attributes. In these attribute-derived networks, the actual task is one of predicting one set of attributes from other attributes, and the network is only used to represent the known attributes in a particular manner adapted for a particular type of machine learning. In another type of network, the network exists independently of the attributes of the nodes, but it is a functional network: for example, in a citation network among academic papers, we know that a link has a specific function in the originating node, since a paper only cites work that is related (be it related and similar or related but different). Networks defined by web pages are also functional networks. In contrast, in social networks with mutually self-declared friendship (which we are interested in), we do not know whether a friendship link corresponds to anything in particular. In general, friendship is not functional, and the reason for a friendship link could be varied and very subtle like common view point on a discussion board, an offline meeting, or a purely visual interest based on a photograph. This paper is in fact about empirically investigating whether friendship links do relate to certain attributes, and we have *a priori* no reason to believe that our mutually self-declared friendship networks behave similarly to attribute-derived and functional networks.

The contributions of our paper are as follows:

- We introduce a new data-set, TravelSite, which contains both a self-declared friendship network, and self-chosen attributes from a finite list defined by the social networking site (Section III).
- We introduce GFHF, a state-of-the-art graph transduction algorithm, as a novel way of testing the relevance of the friendship network and thus the homophily hypothesis. We show that GFHF does make use of the friendship network when homophily is present.
- We show that in our data, homophily correlates with the rate of incidence of an attribute (this is not necessarily

[1]This work was supported by the National Science Foundation's KDD program, a joint program with the National Security Agency. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors only.

the case).
- We use SVM with other features extracted from the data for the task of prediction, and show that other attributes like age, languages spoken etc. are also important (Section IV-C).

We also show that the relative performance on different attributes depends on the task we evaluate. In addition to the standard task of determining whether or not an unlabeled node in the network has an attribute or not (measured using accuracy), we also consider the task of finding all and only the unlabeled nodes in the network which have a certain attribute (measured in f-measure). We find that different attributes profit from prediction from the network depending on task (Section V).

II. Previous Work

The increased use of the internet by an ever growing number of users has resulted in the development of a massive number of online social networks. There exists a lot of information about peoples' networks and attributes on the freely available world-wide web. This has posed many interesting problems for the research community like social network analysis [3], [4], target marketing [5], prediction of future friendships (a task analogous to [6]). We now discuss several papers which are directly relevant to our work.

Lu and Getoor (2003) [7] show that considering unlabeled data in a link network along with the labeled data (i.e., semi-supervised learning) helps. We follow their suggestion and use graph transduction, a different semi-supervised method, though note that machine learning is not the focus of our work – we are primarily interested in empirically analyzing if the underlying friends network helps in the prediction of attributes of people on an online social network. An important difference between our work and [7] (apart from the difference in machine learning algorithm) is the type of data: they use functional networks (citation networks and web page networks), while we use a non-functional friends network.

Lindamood and Kantarcioglu (2008) [8] suggest a modified Naive Bayes approach to predict undisclosed private information on a friendship network (Facebook). Their interests are thus compatible with ours, in that they want to show that a non-functional friends network can help in classifying members of the network for attributes. The main difference to our work is that they investigate only one attribute, and do not vary the amount of training data (in fact, it is not clear how much training data is used). Furthermore, they use modified Bayesian inference, while we use graph transduction and SVMs. Finally, their focus is on developing a methodology to prevent prediction of private information, while our focus is on investigating the empirical relation between the friendship network and different individual attributes.

Like [8], Zheleva and Getoor (2009) [9] also study homophily from the point of view of privacy concerns, investigating several non-functional friends networks and several algorithms. Their goal is thus very similar to ours. The main difference is that they only predict one or two attribute per data set, and the attributes do not belong to the same class. In contrast, the 26 hobbies we are predicting belong to the same class ("interests") and therefore enable us to investigate if the performance of an algorithm remains the same on different attributes belonging to the same class.

The work of He et al. (2006) [2] is perhaps closest to ours: they have a non-functional friends network (LiveJournal) and they are interested in predicting the attributes of people on that friendship network. They take into account only the underlying friendship network while we also incorporate other attribute information in our classifier.

III. Our Corpus

We used WWW::Robot[2], a web traversal engine, to crawl an online travel community, which we will call TravelSite.[3] TravelSite allows members to create mutual friendship links (which exist only when both members have agreed to the link), discuss the places they have visited, declare their hobbies and other attributes like age and languages they speak etc. From the site map of the website, we downloaded a list of $11,000$ users. For each user, we then downloaded their "about me", "my friends" and "pictures" HTML pages. We wrote Perl scripts to extract users' *hometown, languages, gender, age* and *interests* from their "about me" pages, a list of their friends from their "my friends" pages and a list of places they have visited from their "pictures" pages. Specifically, the information we gathered from their website is as follows:

- Friends' Network: This is the mutually self-declared friends network matrix. This feature is encoded as a binary feature where the friends of a person out of 181 people is set to 1 and all the others are 0. All other features are also encoded as binary features in the similar manner.
- Hobbies: Members declare their hobbies by clicking on boxes next to a list of 26 possible hobbies. Thus, each user can have between 0 and 26 hobbies, and we can consider each hobby as a binary feature.
- Countries Visited: Members can upload photographs, for which they must choose a location which includes a country name from a standard set of 264 countries. Thus, every member can declare up to 264 countries (if they upload at least 264 photographs). We also included their home country in this list.
- Languages Spoken: There is a list of 139 languages from which members select a maximum of three languages they speak.
- Age group: The age group is in terms of ranges, example under 20, 20-25, 26-30 etc, from which the user chooses one. There are a total of 12 ranges.

We refer to the countries visited, the languages spoken, and the age group as "personal characteristics". Note that all information is from forms, so there is no free text and no need for processing free text.

[2]http://search.cpan.org/ kventin/WWW-Robot-0.025/lib/WWW/Robot.pm
[3]Please contact us for the data.

Number of nodes in the graph	181
Number of friendship links in the graph	1214
Number of Age groups	12
Number of different languages users speak	139
Number of different countries users may have visited	264
Number of different hobbies a user could have	26
Total number of unique attributes of a node in the graph	441

TABLE I
STATISTICS ABOUT THE DATA

Even though there were many people registered on the website, we found, only a few had more than one friend. In fact, most users have no friends and therefore, the network was very sparse and disconnected. We reduced the set of users under consideration as follows. We looked for people who have over 6 friends, and thus obtained a list of 28 people. We then took these 28 people and all of their friends, giving us 181 people, and included all friendship links between these 181 people. This is the network we use for the experiments reported in this paper. The degree distribution as shown in Figure 1 suggests that the resulting network follows the power rule of Social Networks [10]. The network is smaller than some other networks that have been used in recent studies; however, the fact that we have 26 hobbies which we can potentially investigate, and the fact that they are declared from a form input, allows us to perform experiments which would not be possible on larger data sets.

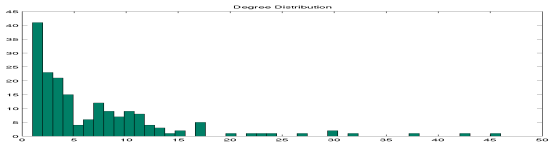

Fig. 1. Distribution of the number of friends (x-axis) against the number of users with that number of friends (y-axis)

IV. EXPERIMENTS AND RESULTS

In this section we first formally define our task in Section IV-A. In Section IV-B, we present a graph transduction approach, Gaussian Field Harmonic Functions (GFHF) [11] to test our research hypothesis: "the self-declared friends network helps in predicting hobbies of people on an online social network." In order to test our hypothesis, we run GFHF using two weight matrices: a) a randomly generated binary weight matrix and b) the self-declared friends network. We then compare the difference in the performance of GFHF on the aforementioned weight matrices over many trials. Finally, we accept or reject our hypothesis based on statistical significance testing. In Section IV-C, we propose the use of prediction by Support Vector Machines (SVM [12]) as "dongles " attached to graph nodes while running GFHF [11]. SVMs are introduced so we can learn from other characteristics and hobbies. The intention here is to test if other attributes help.

A. Formal Problem Statement

Consider a graph $G = (V, E)$. In this graph, $v_i \in V$ are people in the network connected with edges $w_i \in E$. Edges represent a connection between people as declared by them on the network. We refer to G as a "self-declared friends network" or "friends weight matrix" in the paper. If there exists an edge between two people we assign a weight of 1, otherwise 0. Therefore, we treat friendships as binary relationships without the concept of friendship strength. We pair the graph with a function from nodes V to an attribute set $T = C \cup H$ such that each node in the network is associated with a binary feature vector. This vector consists of different classes of attributes: C are personal characteristics and H are the various hobbies a person may have. In this paper we predict only hobbies. Specifically, our task is to predict, for each hobby $h_k \in H$, whether a given person has the hobby or not.

B. Gaussian Field Harmonic Functions

1) Introduction: Graph transduction is a class of algorithms that falls into the category of semi-supervised learning (see [13] for an overview). In a semi-supervised learning paradigm, given a few labeled examples, the classifier uses both labeled and unlabeled examples to predict labels of the unlabeled examples. Specifically, in a graph structure where the labeled and unlabeled examples are connected, [14] discuss that the unlabeled data plays an important role in the correct prediction. The task we are dealing with fits well with the framework of a semi-supervised learning method since we have a set of people whose attributes are known and we are trying to predict the attributes of the remaining people. Since our data is represented in form of a graph, a good choice is GFHF [11]. The algorithm can be seen as a random walk over a graph structure, such that labels are assigned to unlabeled nodes by starting from an unlabeled node and reaching a labeled node based on the probability distribution on the intervening nodes. In our case, these probabilities are binary as we consider the self-declared friendship network as the graph structure on which predictions are to be made.

A set of l nodes are labeled from amongst the $|V|$ nodes of the network. Based on this choice, the network is thought to be divided into a labeled set and an unlabeled set. We assume that the actual labels of the network form a Gaussian field. Point values in a Gaussian field are known to vary harmonically and are uniquely defined. Therefore, solving for the labels of unlabeled nodes is equivalent to finding the unique solution that satisfies the harmonic property. Finding the unique function which satisfies the harmonic property is the same as solving for $f(i)$ in Equation 1 for each unlabeled data point i. Here, u is the number of unlabeled examples and $l + u$ is the total number of examples.

$$f(i) = \frac{1}{D_{ii}} \sum_{j \sim i} w_{ij} f(j) \quad where, i = l+1, ..., l+u \qquad (1)$$

Equation 1 can also be expressed as in Equation 2 where D is the diagonal matrix of W which is the weight matrix.

$$f = D^{-1}Wf \tag{2}$$

We know that,

$$W = \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix} \tag{3}$$

$$f = \begin{bmatrix} f_l \\ f_u \end{bmatrix} \tag{4}$$

Thus, from Equation 2, we get Equation 5 which helps us to calculate the stable harmonic solution of this graph.

$$f_u = (D_{uu} - W_{uu})^{-1} W_{ul} f_l \tag{5}$$

Equation 6 is obtained by substitution $P = D^{-1}W$ in Equation 5.

$$f_u = (I - P_{uu})^{-1} P_{ul} f_l \tag{6}$$

2) How we set up our experiment and results: Our research hypothesis is that there is a correlation between mutual self-declared friendship links in online social networks and attributes listed in the profiles of said friends, presumably because of homophily. [13] finds that the performance of the GFHF algorithm is extremely sensitive to the *correctness* of the weight matrices. Thus, GFHF allows us to test our hypothesis. To accept or reject our research hypothesis, we consider the prediction capability of GFHF using two weight matrices: 1) G_R which is the random weight matrix i.e. connects people on the social network randomly and 2) G_F which is the self-declared friends network. We run GFHF 30 times, each time for a random configuration of n_i number of labeled data points where $n_i \in N = (10, 30, 50, 70, 90)$. We carry out these predictions for all 26 hobbies under consideration. Therefore, for each weight matrix, G_R and G_F we get a corresponding $26 \times 5 \times 30$ matrix, where 26 is the number of hobbies, 5 is the different number of data-points and 30 is the number trials. Figure 2 shows the accuracy of running GFHF with the random matrix (G_R) and with the friends matrix (G_F) for 26 hobbies and across 3 different training set sizes (numbers of labeled data-points; we omit 30 and 70 for lack of space but include them in the averages). The numbers are averages over the 30 trials with the same configuration. The second-to-last column shows the average of difference in accuracy between G_F and G_R across all training set sizes, and the last column shows the difference in accuracy between S_T and G_F, again as average across all training set sizes. For each hobby and a data-point, we then use a Wilcoxon test (for 30 paired accuracies) to see the statistical significance of the results. All differences in % accuracies of paired results in Figure 2 are statistically significant at $p < 0.0001$.

3) Discussion - what we show: Our results show that in most of the cases G_F performs significantly better than G_R which implies that the underlying friends network is in fact important for prediction. For some hobbies (below the horizontal line in Figure 2), the difference in the performance of G_F and G_R is extremely high. These are precisely the hobbies that over 50% of the people in the network have. In fact, $G_F - G_R$ is highly correlated (using the Pearson Correlation Coefficient) with the incidence[4] of the hobbies ($R^2 = 0.74; p < 0.0001$). Does this mean that it is in the nature of the algorithm to perform better when more people have a particular hobby? Or is it not a property of the algorithm but of the data that leads to this correlation? We expect GFHF to exploit homophily and to do better for hobbies that are good factors of homophily and worse for others. To investigate this we find the correlation between homophilous strength of hobbies and hobby incidence. We develop four naive but obvious measures of homophilious strength. The measures attempt to answer the following question: what percentage of people having a particular hobby have over $n\%$ of friends with the same hobby? We use four values for n - 25, 50, 75 and 100. First we calculate the correlation between these measures and incidence. This is found to be high ($R^2 = 0.72, n = 25; R^2 = 0.92, n = 50; R^2 = 0.81, n = 75; R^2 = 0.81, n = 100; p < 0.0001$). This is a crucial property of our data: the incidence of a hobby is highly correlated with the homophilous strength of the hobby. Therefore, as the incidence increases, so does the effect of homophily. We conclude that the fact that GFHF performs much better when the incidence is high simply shows that GFHF is good at exploiting homophilous hobbies. Put differently, GFHF is a good algorithm with which to test homophily.

C. SVM "Dongles" for GFHF

1) Introduction: The focus of this section is to see if the characteristics and other hobbies of users help in the correct prediction of their hobbies. To incorporate the effect of other attributes, use a Support Vector Machine (SVM) along with GFHF.[5] We use two feature sets when using SVMs: in a first set, we use only the set of personal characteristics C. In a second set, we also use all hobbies except the one we are predicting, which gives us $C \cup H \setminus \{h_k\}$. Once we get the predictions for an SVM for varying labeled data size, we use equation 7 to incorporate that into a classifier. Essentially, we have a composite classifier in which GFHF makes use of the friends network and SVM provides support to the predictions by incorporating the other hobbies and characteristics. We call this system S_C when we only use characteristics C and do not use the set $H \setminus \{h_k\}$, while S_T is the system that uses the complete set of attributes $C \cup H \setminus \{h_k\}$.

We implement a scheme as suggested by [11] where they attach "dongles" to each vertex $v_i \in V$ in the graph G. Basically, now $G = (V, E, Q)$ where $|V| = |Q|$ such that every $v_i \in V$ has an attached node labeled with $q_i \in Q$. The GFHF equation them becomes:

$$f_u = (I - (1-\eta)P_{uu})^{-1}((1-\eta)P_{ul}f_l + \eta q_u) \tag{7}$$

[4]By incidence we mean the percentage of people who have a certain hobby. For example, if 2 out of 10 people have a hobby h_1, then the incidence is 20%.

[5]We use Matlab implementation of [15]

H	I	10				50				90				Average	
#	%	G_R	G_F	S_C	S_T	G_R	G_F	S_C	S_T	G_R	G_F	S_C	S_T	G_F–G_R	S_T–G_F
12	6.1	90.8	92.4	94.1	93.6	92.6	92.4	94.6	93.9	92.3	92.3	94.5	94.2	0.3	1.6
22	6.6	92.1	91.7	93.5	93.5	92.6	91.8	93.9	93.4	91.9	90.8	93.4	93.2	-0.8	1.8
11	7.7	89.0	90.9	92.4	92.2	90.2	90.0	92.3	92.0	90.1	90.3	92.3	91.9	0.7	1.5
2	8.3	88.2	90.6	91.8	91.5	89.4	90.6	92.3	91.4	89.0	90.2	92.3	91.6	1.4	0.9
1	9.9	87.9	88.8	90.0	89.9	88.4	87.9	90.1	89.7	88.0	87.2	90.1	89.5	-0.3	1.7
20	12.2	85.7	85.6	88.0	86.2	86.9	85.9	88.3	86.7	85.6	84.9	87.8	85.9	-1.0	1.1
21	12.2	83.8	86.5	87.8	86.3	85.9	86.3	88.0	87.2	85.1	86.0	87.5	87.4	0.8	0.8
10	16.6	77.6	82.1	82.7	81.8	80.2	82.2	83.0	81.0	79.5	79.8	82.9	80.8	1.9	0.0
19	17.1	79.5	81.2	82.9	82.0	80.4	80.9	83.0	82.4	79.7	80.2	83.5	82.6	0.5	1.3
5	20.4	70.8	77.3	75.9	74.9	76.2	76.9	77.1	77.4	75.2	74.7	76.7	77.8	2.0	0.5
9	29.3	62.6	68.9	68.2	68.9	62.1	67.3	68.2	71.6	62.0	64.2	69.2	72.0	4.6	4.1
25	30.4	57.5	65.5	62.8	64.4	61.9	65.9	66.3	66.9	60.7	63.7	67.0	66.8	4.6	1.1
4	31.5	57.9	64.6	66.6	67.5	62.8	63.5	68.2	70.6	63.1	64.3	66.7	70.2	2.9	5.2
24	33.1	54.9	62.6	62.5	63.0	57.7	61.0	64.5	64.9	55.7	58.7	63.8	63.6	4.0	2.8
6	33.7	55.4	62.4	59.8	61.5	60.0	60.9	63.2	67.5	61.7	58.2	61.3	67.7	1.1	5.3
13	35.4	55.7	60.9	62.9	66.6	57.0	58.5	66.2	71.5	56.2	54.6	64.1	72.1	1.5	11.9
23	37.0	55.3	60.4	59.7	59.3	58.0	59.7	58.1	62.4	57.3	59.6	60.4	61.9	2.6	1.6
7	39.2	51.5	58.0	55.6	55.9	51.3	57.8	57.0	57.1	50.1	58.1	58.9	56.6	6.9	-1.0
8	40.9	48.5	48.0	52.2	57.2	50.8	55.8	55.3	62.0	49.9	53.7	56.2	60.3	3.4	7.3
14	41.4	49.2	48.9	51.2	55.8	51.7	51.9	56.0	62.1	50.2	50.6	57.6	64.7	-0.3	10.6
3	47.0	49.8	52.5	49.1	59.4	51.5	56.7	48.8	64.2	52.5	57.6	47.7	66.2	4.4	7.8
15	50.8	49.7	51.1	50.1	55.2	47.9	54.1	46.3	56.4	46.2	53.4	45.8	57.3	5.5	3.8
18	63.5	36.8	56.8	62.5	62.9	36.7	61.8	58.9	64.2	38.1	61.0	58.2	64.8	23.1	3.5
16	67.4	32.7	59.7	64.9	64.7	32.8	63.5	63.4	64.6	33.0	62.0	63.4	64.5	29.5	2.3
17	71.3	29.2	45.8	70.4	70.2	29.0	64.1	68.4	68.5	29.7	62.1	67.6	68.7	30.5	9.3
26	93.4	7.0	7.0	92.8	93.0	6.9	46.7	93.1	93.0	7.7	63.3	92.3	92.2	27.0	58.5

Fig. 2. This table compares four systems: G_R (only GFHF with a random friends network) G_F (only GFHF with friends network), S_C (GFHF with dongles using only C characteristics) and S_T (GFHF with dongles using all attributes except the hobby being predicted). $H\#$ stands for hobby number and $I\%$ for incidence, i.e., the percentage of people who have a particular hobby. In general, S_T performs better than S_C followed by G_F followed by G_R. The second-to-last column shows the average increase (over all data points) in accuracy of G_F over G_R. The last column shows the average increase (over all data points) in accuracy of S_T over G_F. From this table we also observe that as we increase the data, prediction accuracy increases.

where q_u are the predictions of SVM for the unlabeled node. Basically, η is the probability with which the random walk goes to the prediction of SVM. We then run experiments for η varying from 0 to 1 with steps of 0.1. 0 implies pure GFHF and 1 implies pure SVM.

2) How we set up our experiment and results: Our research hypothesis is that the characteristics and the other hobbies help in the prediction of a particular hobby. From the hobbies, we draw one hobby at a time and treat it as the class. We use the remaining hobbies in the feature set which we run through classifiers. In this manner, we ran classification algorithms for 26 different hobbies. Figure 2 shows results for the comparison of three systems: G_F (only GFHF with friends network), S_C (GFHF with dongles using only C characteristics) and S_T (GFHF with dongles using all attributes except the hobby being predicted). In general, S_T performs better than S_C, which performs better than G_F. The last column shows the average increase (over all data points) in accuracy of S_T over G_F. From this table we also observe that as we increase the data, prediction accuracy increases for the SVM (again with some exceptions).

3) Discussion - what we show: Figure 2 shows that our research hypothesis holds: using additional attributes helps the prediction. Interestingly, in the results over all hobbies, we observed that the GFHF and SVM don't really complement each other. In fact, around $\eta = 0.4$ the accuracy stabilizes and equals the that of pure SVM prediction with full attribute list ($\eta = 1$). This motivates the need for new algorithms that have an additive affect or incorporate other attribute information into GFHF in a meaningful way. However, it is also possible that homophily is indeed in action, and access to the 25 other hobbies provides the same information about the structure of the friends network that using the friends network directly would provide. This issue requires further study.

V. Finding All People with a Specific Hobby

So far, we have been examining the following task: given a social network of people, for some of whom we know whether or not they have a particular hobby h, can we predict whether or not the others have hobby h? We use accuracy to evaluate this task. However, one might also be interested in a different task: given a social network of people, for some of whom we

know whether or not they have a particular hobby h, can we find all people who have the hobby h? This task is properly evaluated using recall and precision, and their combination, f-measure. It is important to understand that the two measures, accuracy and f-measure, measure the performance of a system on different tasks, so one has to be clear about the tasks.

We evaluated three systems on f-measure: as a baseline, the GFHF-based system with a random matrix (G_R); the GFHF-based system with the actual friends matrix (G_F); and the system that includes an SVM, using all features, including the other hobbies (S_T). The results are shown graphically in Figure 3, where we report the f-measure on unlabeled data, averaged over training on 10, 30, 50, 70, and 90 labeled nodes. As is the case when we use accuracy as a measure, we find that S_T always performs as well or (usually) outperforms both G_R and G_F. However, the relation between the baseline G_R and the system using only the friends network (G_F) is rather different. As we can see form Figure 3, there are quite a few hobbies for which the friends network does not provide any useful information.

Why is this? In Figure 3, we have arranged the hobbies by increasing incidence of the hobby in our social networks (the relative can be seen in the second column of Figure 2). We see that the friends network does not consistently help over a random network if the hobby has a relative incidence of 41% or less (there are 10 hobbies up to hobby 14 for which the random network outperforms the friends network), while at 47% and above (starting with hobby 3), the friends network consistently outperforms the random network. In contrast, the accuracy data contrasting G_R and G_F shows that using the friends network helps, except with three very rare hobbies (incidence less than 13%), and one outlier. F-measure, in contrast to accuracy, does not take the number of correctly predicted negatives into account, and we conclude from the data presented here that G_F, for hobbies with relative incidence less that 45%, is better at identifying people without the hobby than with the hobby. It is only when a hobby is sufficiently frequent that GFHF succeeds in identifying positive instances using the social network.

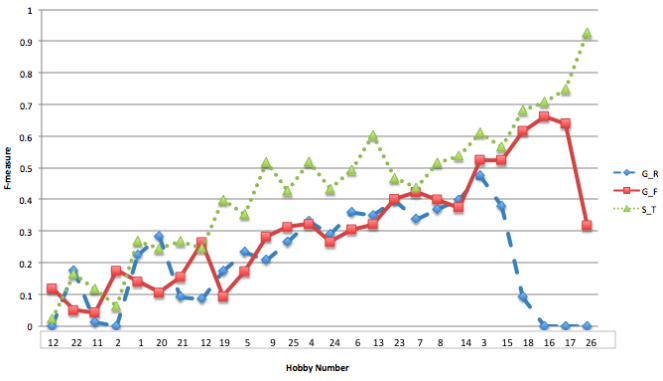

Fig. 3. Using f-measure to evaluate the performance of G_R (dashed line), G_F (solid line), S_T (dotted line)

VI. Conclusion and Future Work

We have introduced a new data-set, TravelSite, which contains both a self-declared friendship network, and self-chosen characteristics and hobbies from a finite list defined by the social networking site. We have proposed GFHF, a state-of-the-art graph transduction algorithm, as a novel way of testing the relevance of the friendship network and thus of the homophily hypothesis. We have shown that the underlying self-declared friendship network allows us to predict hobbies of people in our online social network, and the conrtribution of the friendship betwork increases with the incidence of the hobby we are predicting. We have also used SVMs with the personal characteristics and the other hobbies as features, and show that other attributes are also important. Finally, we have shown that if we switch our task to one of finding all and only the users who have a given hobby, the friends network is useful only for the more frequent hobbies.

References

[1] M. Mcpherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, vol. 27, pp. 415–444, 2001. [Online]. Available: http://dx.doi.org/10.2307/2678628

[2] J. He, W. Chu, and V. Liu, "Inferring privacy information from social networks," in *Intelligence and Security Informatics*, Mehrotra, Ed., vol. LNCS, no. 3975, 2006.

[3] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks." *Nature*, vol. 393, no. 6684, pp. 440–442, June 1998. [Online]. Available: http://dx.doi.org/10.1038/30918

[4] P. Domingos and M. Richardson, "Mining the network value of customers," in *Seventh International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA: ACM Press, 2001.

[5] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2003, pp. 137–146.

[6] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein, "A bayesian networks approach for predicting protein-protein interactions from genomic data." *Science*, vol. 302, no. 5644, pp. 449–453, October 2003. [Online]. Available: http://dx.doi.org/10.1126/science.1087361

[7] Q. Lu and L. Getoor, "Link-based classification," in *ICML*, T. Fawcett, N. Mishra, T. Fawcett, and N. Mishra, Eds. AAAI Press, 2003, pp. 496–503.

[8] J. Lindamood and M. Kantarcioglu, "Inferring private information using social network data," University of Texas at Dallas, Dallax, TX, Tech. Rep., 2008.

[9] E. Zheleva and L. Getoor, "To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles," in *18th International World Wide Web conference (WWW)*, April 2009.

[10] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman, "Search in power-law networks," *Phys. Rev. E*, vol. 64, no. 4, p. 046135, Sep 2001.

[11] X. Z. Zhuxj, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *In ICML*, 2003, pp. 912–919.

[12] V. N. Vapnik, *The nature of Statistical Learning Theory.* Springer, 1995.

[13] X. Zhu, "Semi-supervised learning literature survey," Computer Sciences, University of Wisconsin-Madison, Tech. Rep., 2005.

[14] M.-F. Balcan, A. Blum, P. P. Choi, J. Lafferty, B. Pantano, M. R. Rwebangira, and X. Zhu, "Person identification in webcam images: An application of semi-supervised learning." in *Proceedings of International Conf. on Machine Learning Workshop on Learning from Partially Classified Training Data*, 2005, pp. 1–9.

[15] S. Gunn, "MATLAB support vector machine toolbox," (Internet), Mar. 1998. [Online]. Available: http://www.isis.ecs.soton.ac.uk/isystems/kernel