Predicting Collective Identities in Social Media Graphs

Roghayeh Feizi
Dept. of Computer Science
University of Rochester
Rochester, NY 14627
rfeizi@cs.rochester.edu

Yijun Huang
Dept. of Computer Science
University of Rochester
Rochester, NY 14627
huangyj0@gmail.com

Ji Liu
Dept. of Computer Science
University of Rochester
Rochester, NY 14627
jliu@cs.rochester.edu

Christopher Homan
Dept. of Computer Science
Rochester Institute of Tech.
Rochester, NY 14623
cmh@cs.rit.edu

Vincent Silenzio
Dept. of Psychiatry
University of Rochester
Rochester, NY 14627

Henry Kautz
Dept. of Computer Science
University of Rochester
Rochester, NY 14627
henry.kautz@gmail.com

ABSTRACT

We provide a robust framework for predicting membership in groups of collective identity. We show that, with a dense online social network graph that corresponds to a geographic locale, we can predict individuals who self-identify as gay or lesbian, or as teenagers, with a very small amount of training data and using strictly the tie structure of the graph.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: [On-line Information ServicesâĂŤWeb-based services]; I.2.6 [**Artificial Intelligence**]: [Learning]

General Terms

Algorithms, Design, Experimentation

Keywords

Social Media, Social Network Analysis, Classification, Label Propagation

1. INTRODUCTION

Both age and sexual orientation are examples of traits that are inherent of the individual (in the sense of how one experiences oneself and interacts with others), and that support the development of homophily or assortativity within a socially interactive network [18]. This facet of social network dynamics can be exploited to both predict network and vertex-level features.

Collective identity is the sense of belonging to a group of individuals who share certain personal traits, e.g., age, physical appearance, political affiliation, or sexual orientation. It is a central component of much sociological research [2, 4]. There are numerous concrete reasons to study this phenomenon. As one example, many groups based on collective identity face distinct and pressing health challenges. For instance, alcoholism and drug abuse often begins in adolescence [7]. Lesbian, gay, bisexual, transgendered, and questioning (LGBTQ) individuals are between two and seven times as likely as their non-LGBTQ counterparts to express suicide-related thoughts or behaviors [5, 10].

The variation in suicide rate estimates in LGBTQ individuals illustrates a significant challenge to studying collective identity: many such groups are relatively small and function to protect themselves from the scrutiny of and stigmatization by the larger population, sometimes making them hard to discover or even estimate their size. There are also ethical concerns about the risk of exposure that studying such hidden groups can incur [16].

We investigate a variety of machine learning approaches for determining the collective identities of individuals in an online social network. We show that with a very small amount of training data (3492 labeled users with only 377 positive labels in a social network of 251,316 users in one case) and nothing more than the publicly-expressed social ties within this dense network, we can predict whether individuals fall into one of two different, important collective identities: (1) adolescents and (2) lesbian and gay individuals (the latter of which additionally presents a significant class imbalance problem), at success rates much higher than reasonable baselines.

A novel aspect of our work is how we collect the data. Most prior work on predicting collective identities in online media first collects gold-standard labeled individuals from the medium, then collects the immediate social neighbors of each individual. The result is a series of small social graphs, with no necessary connections between them. Given that the social network is itself evidence of the complex processes by which people organize themselves around collective identities, there would seem to be a great deal of useful information in the space between such neighborhood graphs.

In contrast, we first collect a dense portion of the social graph that includes as many users as possible from a large underlying population. We then extract gold-standard exemplars from this functional, comprehensive slice of the social graph. This provides us an *in situ* social context that helps us discriminate members of the identity-based groups of interest from the general population and provides a a

much stronger signal than those provided by other, seemingly more obvious signals of collective identity, such as a language used or celebrities followed. In effect, the social network serves as the output of a collective, assortative, human computational process for determining group identity that a machine learning algorithm might otherwise need to perform.

To our knowledge, this is the first research into the use of machine learning to predict sexual orientation. Both age and sexual orientation illustrate a number of interesting properties from a collective identity perspective.

For one, the trait(s) used to determine such identities often fall individually along a spectrum. For age this is self-evident. For sexual orientation, think of the Kinsey scale [16]. Yet collective identities create a discretized perception of the trait by including those with "enough" of it and excluding those without. This process of discretization has its reasons, e.g., to provide coalitional power to otherwise marginalized people, but also creates stress on its members who fall near the threshold points of the spectrum, e.g., bisexual individuals have an even higher rate of suicide than their gay or lesbian peers. One of the aspects of the machine learning algorithms we study is whether it is better to model internally the believed group membership of on an individual as a discrete or continuous variable.

This process of discretization itself is governed by two competing forces: *selection* (or *homophily*), i.e., the tendency of individuals with similar traits to associate with one another, and *influence*, or the tendency for the traits of associated individuals to become more similar over time. For certain types of collective identity, such as political affiliation, it is challenging to disentangle these forces. The two identities we study, however, are very clearly based on traits that are, on the timescales we study at least, primarily selective forces on the network, rather than the product of social influence.

2. RELATED WORK

Latané and L'Herrou [9] pioneered the use of computer models (via simulation) to study how network structure serves to reinforce the opinions and beliefs of strongly connected components, shedding light on the social functions that collective identity plays. DiFonzo et al. [4] used computer-mediated human experiments in which respondents were drawn equally from Republicans and Democrats to study the specifically how collective identity moderates the impact of negative in-group rumors, especially when the group have stronger intra-group connectivity.

Rao et al. [15] classify users into age, gender, and origin categories based on language variation. They use a set of socio-linguistic features, such as n-grams of tweet messages. They achieve an accuracy of 74.11 when classifying users into those younger or older than 30. They use network features, but only at a descriptive level, i.e., they count the number of links a user has, but do not model the edges as traces of social activity, as we do. We, by contrast use the network as system of relational constraints on the attributes of unobserved individuals and use those constraints to infer those attributes..

Al Zamal et al. [1] study the Twitter users' gender, age, and political affiliation. They use the content of the users' tweets as well as those of their immediate neighbors' along with several statistical features obtained from profiles as a feature vector in a supervised approach. We by contrast first draw from a tightly connected network and then discover members of groups of interest from the members of this network only. This allows us to study the effects of network structure beyond the friendship level.

Pennacchiotti and Pope [13] classify in separate experiments users along the collective identities of political affiliation (Democratic or Republican), ethnicity (African-American or not), and whether or not they are Starbucks fans. They gather gold-standard labels by scraping Twitter directories for political affiliation, searching for explicit mention of African-American ethnicity, or gathering those who follow Starbucks on Twitter. They they study a variety of linguistic and user profile features. Finally they use a simple one-round relabeling process based on the labels of each subjects friends. Our work differs from their in several respects. First, and most importantly, they gather their data by drawing from Twitter users as a compartmental frame, i.e., without regard to the underlying network. Only after they have their sample populations do they gather the ego networks of each sample member. Thus, the resulting graphs has little structural integrity, and network effects can only extend to the neighborhood level (and, in fact, that is as far as their models use network effects). We, by contrast, select the network first, by drawing user from a socially-cohesive locale, and only then search for labeled data. Our goal is in fact to study whether a functioning social network encodes within its structure strong indications of collective identity. Second, most of the nodes in our dataset are completely unlabeled; they serve only to transmit patterns of association.

Peersman et al. [12] study age and gender based on text messages in Netlog social network. They apply text classification over Dutch chat messages drawn from a balanced dataset using a set of token and bag of words features. They show that for classifying the users into young and old categories, certain words and emotions act as important features. The main difference between our approach and theirs is using the social graph and connections between users and studying the whole community.

Although the details regarding methods and analyses have not been extensively documented, Stephens-Davidowitz [17] has published the results of his research efforts to attempt to quantify the number of gay men in the U.S. using data derived from surveys, social networks, and search queries for pornographic materials.

3. METHODS

3.1 Data

For the past eighteen months, we gathered Twitter data from the metropolitan Rochester, NY area, resulting in approximately 15 million tweets. Each week we downloaded the portion of the Twitter followers graph consisting of all users whose tweets we collect, as well as all of their followers and "friends" (Twitter terminology for those who follow them).

To address Twitter-imposed rate limits, we run multiple parallel instances of the Twitter streaming API. We specify a bounding box that includes Rochester, NY and the 15 counties surrounding it. Twitter then streams a sample of tweets that: (a) were geotagged within this box OR (b) were made by those who declare their home to be in a region that intersects the bounded region. By varying the size of the

bounding box and observing the download rate, we validated that we have all tweets satisfying (a) or (b).

Since we are primarily interested in network structure as a trace of the social activity between peers who share a collective identity, we only consider the largest connected component of this graph after we first eliminates all edges from the social graph that are not reciprocated (since non-reciprocated edges indicate a non-peer-like relationship). This results in a network of 251,316 users.

To gather training, test, and evaluation data on sexual orientation, we filtered the profiles from our Twitter social graph. Via manual inspection and trial-and-error, we labeled as "gay" (resp., "straight") users those whose lemmatized profile descriptions case-insensitive matched any of the keywords: "homosexual", "homo", "lesbian", "dike", or "gay" (resp., "straight", "heterosexual", "husband", or "wife"). We then inspected this raw set of filtered data and removed users whose profiles did not clearly indicated sexual orientation. This yielded 4658 users total, 377 of whom were labeled "Gay" and the rest were labeled "Straight." We held out 1166 of these users (of whom 95 were labeled "Gay") for final evaluation, and trained our algorithms on the rest. The subgraph induced by the entire set of labeled set of users on this network has 3100 distinct connected components, the largest of which has 1231 nodes. This subgraph exhibits strong homophily. Of the edges in the subgraph, the likelihood of an alter being gay is 62% if the ego is gay and 2% if the ego is straight.

Labeling users by age proved to be more challenging. We filtered users whose profiles or tweets explicitly mention age or birthdays, along with phrases indicative of adulthood such as "getting married", "my husband"' and "my wife," and whether user frequently tweeting near high schools. This yielded 6,388 users after pre-processing. Since age was still not as clear-cut as the sexual orientation data, we created an Amazon Mechanical Turk Human Intelligence Task (HIT) to have crowd workers take another look at the data. Each Twitter user was evaluated by three different AMT workers, who were provided the profile picture of the user, the tweets that matched the filtering criteria we used, and a link to the profile was also provided.

Preliminary tests on this age-labeled data yielded weak results. Part of the reason for this, we hypothesized, was that Twitter's location-based filters err heavily on the side of inclusiveness. In particular, criterion (b) allows for many many whose declared home lies far outside of Rochester, leading to a social that is not well connected or representative of a natural social network. To correct for this, we filtered out all users from the 251,316-sized set, including only those who sent a geotagged tweet from within the smallest bounding box containing Monroe county. After retaining only the largest connected component on reciprocated links, this reduced the size of the social network to 6,388, of which 235 adult users and 121 teenagers were also in our labeled set. The subgraph induced by these labels contains 148 connected components, the largest of which contains 189 users. This subgraph, compared to the labeled sexual orientation subgraph, exhibited a much milder degree of homophily. The likelihood of an alter being a teenager was 58.36% if the ego was a teen and 32.99% if the ego was an adult.

This surprised us. We expected homophily to be quite pronounced in this population, since teenagers are so seemly preoccupied with establishing social independence from their parents and other authority figures. On reflection, however, these milder levels of homophily made sense, especially when compared to lesbian and gay individuals; teens still depend highly on adults. On the other hand, lesbian and gay individual often face social exclusion and often have to rely on ingroup ties for critical emotional support [8].

We held out 90 of the users from the age-related labeled set (containing 31 teenagers) for final evaluation and used the remaining for training.

We tried using this smaller set of 6,388 users with the sexual-orientation-labeled data, but only three of the users labeled "Gay" remained, so we performed our sexual orientation tests on the larger 251,316-user set.

3.2 Features

We regard all labeled and unlabeled users nodes in a graph $G = (V, E)$, with edge weights representing is the degree of similarity between a pair of users according the following measures.

FriendEdge This is just the adjacency matrix of the graph.

$$FriendEdge(i,j) = w_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ follow each other,} \\ 0 & \text{otherwise.} \end{cases}$$

mFriends We define the relative size of the overlap between the neighbors(friends) of users to be a similarity measure between them.

$$w_{ij} = \frac{|N(i) \cap N(j)| + FriendEdge(i,j)}{\min\{|N(i)|, |N(j)|\} + FriendEdge(i,j)}$$

Where N(i) is the set of neighbors of $user_i$

Linguistic content (text). For finding the similarity of users based on their tweets we use an special tokenizer to capture words as well as emoticons like <3 , :) etc.
We then define w_{ij} as the cosine similarity between the vectors of token frequencies in the tweets of users i and j.

Common celebrities followed (gceleb). We define the celebrities to be the users who are followed by more than thousands of users in the dataset (in our experiment we set the threshold to be 200k). The similarity matrix based on this feature is defined by $w_{ij} = \frac{|C(i) \cap C(j)|}{min\{|C(i)|, |C(j)|\}}$, where $C(i)$ is the set of celebrities followed by user i

Common local celebrities followed (lceleb). We find the celebrities that are specific to Rochester,NY. This might include, some high schools, or organizations or famous users who have tweets in Rochester based on the geographical information attached to the tweets.We define a local celebrity to be a twitter user that has more than 5000 followers in our experiments. Then we define the same similarity measure mentioned above for celebrities, for this set of celebrities as well.

3.3 Quadratic Cost Function

The problem of semi-supervised learning on the graph G consists of finding a labeling of the graph that is consistent with both the initial (incomplete) labeling and the geometry of the data that is formed by the graph structure (edges and weights). Given a labeling $\hat{Y} = (\hat{Y}_l, \hat{Y}_u) \in [-1, 1]^n$, we

measure agreement with the initial labeling $Y = (Y_l, Y_u) \in [-1,1]^n$ by

$$\sum_{i=1}^{n} u_i(\hat{y}_i - y_i)^2 = U \cdot \|\hat{Y}_l - Y_l\|^2$$

where U is a vector of weights. In our case, these weights are infinite, i.e., we impose the hard constraint that the initially labeled nodes never change.

To penalize labelings that correlate poorly with the edge weights, we add to the loss function

$$\begin{aligned} L(\hat{Y}) &= \sum_{i,j=1}^{n} w_{ij}(\hat{y}_i - \hat{y_j})^2 \\ &= \sum_{i=1}^{n} {\hat{y}_i}^2 \sum_{j=1}^{n}(w_{ij} + w_{ji}) - 2\sum_{i,j=1}^{n} w_{ij}\hat{y}_i\hat{y}_j \\ &= \hat{Y}^T(D - W)\hat{Y} = \hat{Y}^T L\hat{Y} \end{aligned} \tag{1}$$

where $L = D - W$ is the unnormalized graph Laplacian [19] and we view this as a matrix W, where w_{ij} the weight of edge between i and j.

3.4 Optimization Framework

When $\hat{Y}$ is allowed to take on real values (perhaps representing the degree of confidence we have that an individual belongs to one group or another), the system above is convex and there are a number of feasible approaches for computing a value Y^* that minimizes the quadratic cost functions. However, if we restrict $\hat{Y}$ to discrete values in $\{-1,1\}^n$, it is NP-hard to find the value of Y^* that optimizes the function. The methods we used to approximate Y^* consist of an initialization and an iterative phase, and these can be decoupled and recombined to form a number of different methods. Each approach can be further divided into those that restrict $\hat{Y}$ to $\{-1,1\}^n$ at every step and relaxation methods, which treat it as a member of $[-1,1]^n$ until the end, when it is rounded to some value in $\{-1,1\}^n$.

3.5 Continuous approaches

(Relaxed) label propagation (RLP). In order to minimize the quadratic criterion, we can compute its gradient with respect to $\hat{Y}_u$. This is equivalent to forcing $\hat{Y}_l = Y_l$ and minimizing $\hat{Y}^T L\hat{Y}$. So we can write $\hat{Y} = (Y_l, \hat{Y}_u)$ (i.e. $\hat{Y}_l = Y_l$) and $L = \begin{bmatrix} L_{ll} & L_{lu} \\ L_{ul} & L_{uu} \end{bmatrix}$. The goal of minimizing $\hat{Y}^T L\hat{Y}$ with respect to $\hat{Y}_u$ leads to

$$L_{ul}Y_l + L_{uu}\hat{Y}_u = 0 \Rightarrow \hat{Y}_u = -L_{uu}^{-1}L_{ul}Y_l.$$

The idea in the iterative algorithm is to propagate labels on the graph. Starting with the labeled nodes (nodes 1 to l labeled with 1 or -1), each node starts to propagate its label to its neighbors. This process continues until convergence. W is the similarity matrix and D is a diagonal matrix defined by $D_{ii} = \sum_i^n Wij$.

In this algorithm $\hat{Y}_l$ is constrained to be equal to Y_l. The minimization problem in section 5, are shown to be equivalent to label propagation algorithm presented in this section.

Algorithm 1 (Relaxed) label propagation [19]

1: Compute similarity matrix W
2: Compute the diagonal degree matrix D by $D_{ii} \leftarrow \sum_j W_{ij}$
3: **while** not converged **do**
4: $\quad \hat{Y}^{(t+1)} \leftarrow D^{-1}W\hat{Y}^{(t)}$
5: $\quad \hat{Y}_l^{(t+1)} \leftarrow Y_l$
6: Label point y_i by the sign of $\hat{y}_i^{\infty}$

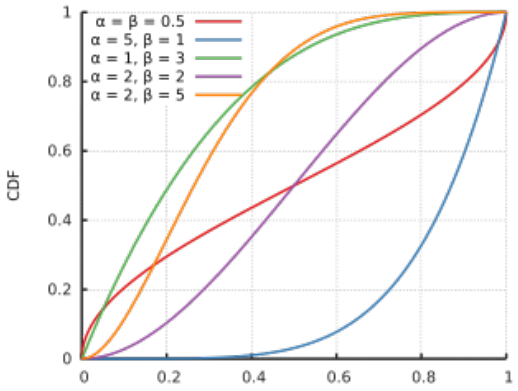

Figure 1: The cumulative distribution function of the beta distribution.

Beta-reweighted label propagation (BRLP) The similarity matrix is the key factor to influence the performance of the label propagation algorithm. Our similarity matrices are constructed from the friendship and the mutual friends. We combine two similarity matrices to improve the performance. Let α be a positive number between 0 and 1 and $W = tW^{(F)} + (1-t)W^{(M)}$.

We apply a monotonically increasing function to map the original "W_{ij}" to new "W'_{ij}", which leads to a reweighted label propagation objective

$$\sum_{i=1}^{n} W'_{ij}(y_i - y_j)^2$$

where $W'_{ij} = F(W_{ij}; \alpha, \beta)$ and $F(\cdot; \alpha, \beta)$ is the cumulative density distribution function (CDF) of the Beta distribution with parameters α and β (see Figure 1). We set $\alpha = \beta$ to make the CDF curve rotationally symmetric at $(\frac{1}{2}, \frac{1}{2})$. See Figure 1.

3.6 Discrete Approaches

Discrete label propagation (DLP). This is the same as relaxed label propagation, except that line 4 reads

$$Y^{(t+1)} \leftarrow T_\beta(D^{-1}W\hat{Y}^{(t)})$$

where T_β is the $\{-1,1\}$ threshold function defined on on $X = (x_i) \in [-1,1]^n$ as

$$(T_\beta(X))_i = \begin{cases} 1 & \text{if } x_i > \beta \\ -1 & \text{otherwise.} \end{cases}$$

Greedy random walks. In this discrete approach, in each round, let $\hat{Y}$ be the current assignments to the nodes. We construct a new assignment $\hat{Y}'$ by choosing k nodes uniformly at random. Then, with probability δ we set inde-

pendently each node to 1. Otherwise, we set it to -1. If $L(\hat{Y}') < (\hat{Y})$ then replace $\hat{Y}$ with $\hat{Y}'$

Metropolis-Hastings random walks (MHRW) [6]. This discrete approach is a form of Markov chain Monte Carlo random walk. In each round, we construct $\hat{Y}'$ as in the greedy way however, here we replace $\hat{Y}$ with $\hat{Y}'$ with probability

$$\min\{e^{L(\hat{Y})-L(\hat{Y}')} \cdot P(\hat{Y}|\hat{Y}')/P(\hat{Y}'|\hat{Y}), 1\},$$

where $P(X|Z)$ is the probability of choosing X as the candidate next assignment state (i.e., $\hat{Y}'$) as the next step, given that Z is the current assignment state. Note that if $P(\hat{Y}|\hat{Y}') = P(\hat{Y}'|\hat{Y})$ then this process always accepts $\hat{Y}'$ when $L(\hat{Y}') < L(\hat{Y})$, i.e., it makes a greedy choice. Otherwise, with some probability makes the switch even though it is not locally optimal.
When we choose each successive $\hat{Y}$ this way, as the length of the walk approaches infinity the likelihood that the last state is $\hat{Y}$ is proportional to $e^{-L(\hat{Y})}$.

3.7 Initialization

The discrete iterative approaches in principle converge to a value or values that are have well-known and potentially useful properties. However, the rate at which convergence may occur can be infeasibly long. It is thus common to assign initial values to the model in a principled fashion. We consider several different approaches.

Relaxed greedy label initialization. In this case we successively assign values to unlabeled nodes, where at each step with choose the unlabeled node that, intuitively speaking, we know (or believe) the most about. See Algorithm 2.

Algorithm 2 Relaxed greedy label initialization (RLI)

1: Let P be a priority queue containing all unlabeled nodes.
2: **while** $P \neq \emptyset$ **do**
3: Pop $i \in P$ that maximizes $\sum_{j \in N(i):j \text{ is labeled}} w_{ij}/Z_i$, where $Z_i = \sum_{j \in N(i)} w_{ij}$.
4: $\hat{y}_i^{(t+1)} \leftarrow \sum_{j \in N(i):j \text{ is labeled}} w_{ij} \cdot \hat{y}_j^{(t)}/Z_i$
5: Label point y_i by the sign of $\hat{y}_i^{\infty}$

Discrete greedy label initialization (DLI) This is the same as the relaxed algorithm, except that line 4 is.

$$y_i^{(t+1)} \leftarrow \begin{cases} 1 & \text{if } \sum_{j \in N(i):j \text{ is labeled}} w_{ij} \cdot y_j/Z_i > \gamma \\ -1 & \text{otherwise.} \end{cases}$$

4. EXPERIMENTAL RESULTS

In most cases, relaxed or beta label propagation was the best performing algorithm. Results on the features were less conclusive, however a number of the features and algorithms yielded relatively weak results. For instance, the greedy and Metropolis-Hastings random walks yielded results that were only marginally better than initialization algorithms alone.

4.1 Sexual Orientation

Table 1 compares the performance of a number of the algorithms using the FriendEdge feature only. We compare their performance to that of the baseline approach of always choosing the most frequent class. BRLP is by far the best performer of these algorithms.

After applying the label propagation algorithm, the best testing accuracy was 93.48%. The precision-recall and receiver-operator characteristic (ROC) curves are shown in Figures 2 and 3.

	Accuracy	Precision	Recall	F-score
Baseline	91.7%			
BRLP	93.05%	67.50%	28.42%	0.4
RLI-DLP	92.9%	59.4%	21.6%	0.32
DLI-DLP	93.0%	77.4%	11.64%	0.20
RLI	92.6%	54.5%	20.5%	0.30
DLI	90.8%	28.6%	13.6%	0.18

Table 1: Sexual orientation results using the FriendEdge feature and the learning algorithms: BRLP and relaxed and discrete greedy initialization, with either label propagation or discrete label propagation. Here, gay and lesbian individuals are the positive class.

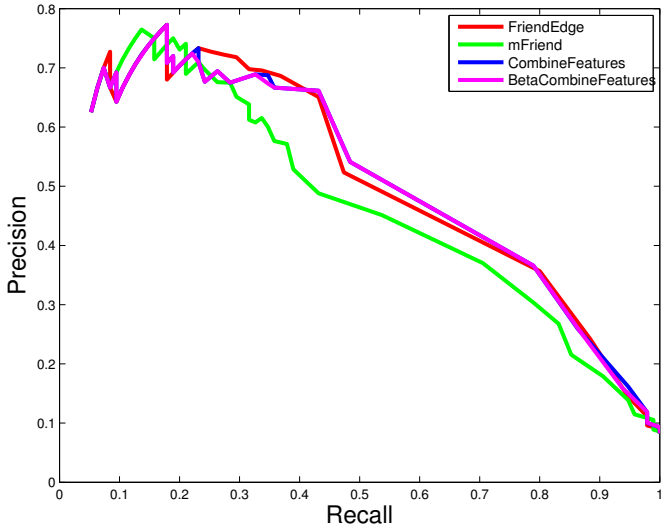

Figure 2: The precision-recall curve for BRLP and RLP on sexual orientation data. "mFriend" and "FriendEdge", denote the RLP algorithm using different similarity matrices and their names indicate the used similarity matrices respectively. "CombineFeatures" uses RLP algorithm with the best combination of all similarity matrices. "BetaCombineFeatures" denotes the BRLP approach with the best combination of all similarity matrices and the best choice of parameter β.

4.2 The age dataset

Table 2 compares the performance of a number of the algorithms using the FriendEdge feature only. We compare their performance to that of the baseline approach of always choosing the most frequent class. BRLP is, again, the best performer.

The precision-recall and (ROC) curves are shown in Figures 4 and 5.

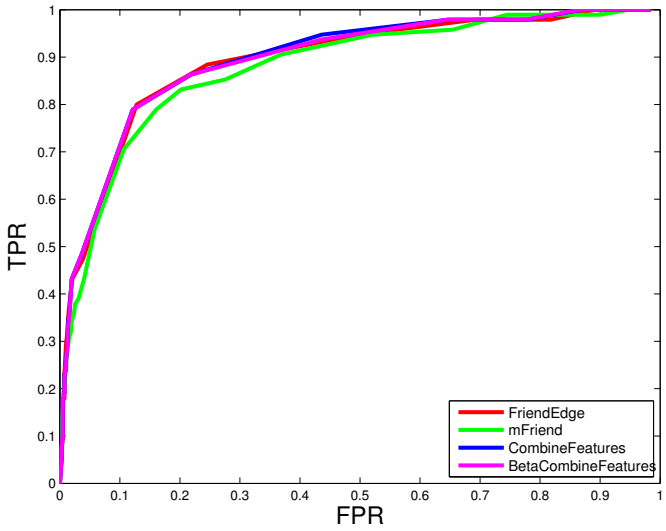

Figure 3: The roc curve for BRLP and RLP on sexual orientation data. Please refer to the caption of Table 2 for the names of all approaches.

	Acc.	Prec.	Recall	F-score
Baseline	65.56%			
mFriends	76.67%	67.86%	61.29%	0.64
FriendEdge	73.33%	70.59%	38.71%	0.5
gceleb	74.44%	83.33%	32.26%	0.47
lceleb	65.56%	0%	100%	0
text	66.67%	60.00%	9.68%	0.17
CombineFeatures	78.89%	70.00%	67.74%	0.69
BetaCombineFeatures	80.00%	70.97%	70.97%	0.71

Table 2: Results of LP on age data using different features. "mFriend", "FriendEdge", "gceleb", "lceleb", and "text" denote the RLP algorithm using different similarity matrices and their names indicate the used similarity matrices respectively. "CombineFeatures" uses RLP algorithm with the best combination of all similarity matrices. "BetaCombineFeatures" denotes the BRLP approach with the best combination of all similarity matrices and the best choice of parameter β.

	Acc.	Prec.	Recall	F-score
Baseline	91.7%			
mFriend	93.05%	67.50%	28.42%	0.4
FriendEdge	93.48%	65.08%	43.16%	0.519
CombineFeatures	93.57%	66.13%	43.16%	0.522
BetaCombineFeatures	93.57%	66.13%	43.16%	0.522

Table 3: Results of LP on sexual orientation data using different features and BRLP. Please refer to the caption of Table 2 for the names of all approaches.

5. CONCLUSION

This paper studied a variety of approaches for exploiting signals of collective identity encoded in the transitive structure of social media networks. Our algorithms were able to perform much better than baseline approaches, even when the vast majority of the data points are unlabeled.

We are acutely aware of the ethics of doing research to detect or discover hidden or otherwise vulnerable communities. For instance, despite the obvious pace of social progress in recent years, sexual orientation remains an intensely private concern for many LGBTQ individuals. Revealing an individual's sexual orientation can still potentially lead to very negative social or economic consequences, to say nothing of the potential for homophobic violence or victimization for at least some of these individuals, particularly youth [3, 11].

We feel that it is nonetheless important to disseminate our work publicly. We believe that having a better understanding of the membership and behaviors of underrepresented groups can lead to public policy benefits for these groups that may far outweigh the risks. Putting these results forward also helps to raise awareness of the extent to which people may reveal hidden information about themselves, simply through their associations.

	Accuracy	Precision	Recall	F-score
Baseline	65.56%%			
BRLP	76.67%	67.86%	61.29%	0.64
RLI-DLP	65.6%	65.6%	100%	0.79
DLI-DLP	65.6%	65.6%	100%	0.79
RLI	70.0%	75.8%	79.7%	0.78
DLI	65.6%	65.6%	100%	0.79

Table 4: Age results using the FriendEdge feature and the learning algorithms: BRLP and relaxed and discrete greedy initialization, with either label propagation or discrete label propagation.

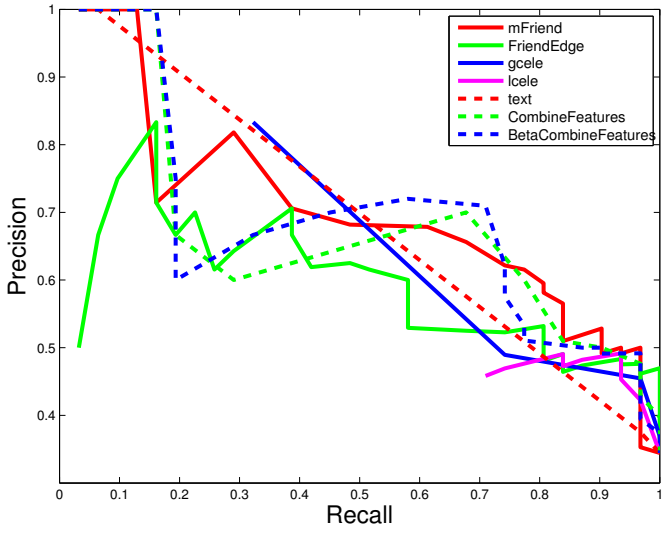

Figure 4: The precision-recall curve for age data. Please refer to the caption of Table 2 for the names of all approaches.

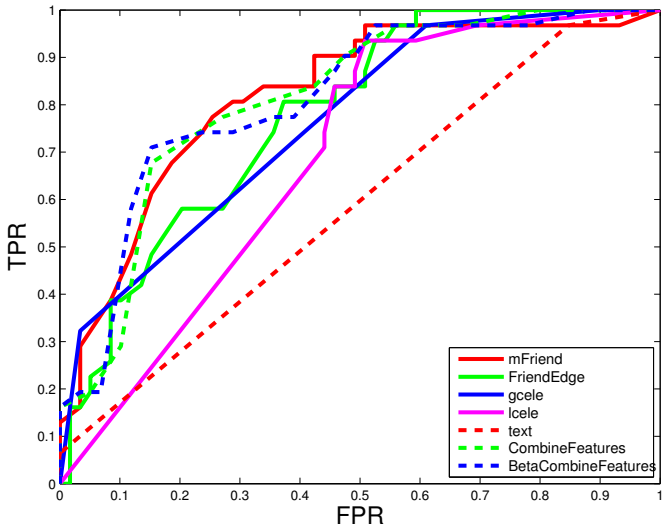

Figure 5: The roc curve for age data. Please refer to the caption of Table 2 for the names of all approaches.

	Accuracy	Precision	Recall
Baseline	92.4%		
Relaxed LP	92.4%	50.0%	11.5%
Relaxed LI	92.8%	80.3%	00.2%
Discrete LI	93.0%	57.1%	27.6%

Table 5: Results of relaxed label propagation and relaxed and discrete greedy initialization, with either label propagation or discrete label propagation.

6. REFERENCES

[1] F. Al Zamal, W. Liu, and D. Ruths. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. *ICWSM*, 270, 2012.

[2] J. E. Côté and C. G. Levine. *Identity, formation, agency, and culture: A social psychological synthesis*. Psychology Press, 2014.

[3] A. R. D'Augelli and A. H. Grossman. Researching Lesbian, Gay, and Bisexual Youth: Conceptual, Practical, and Ethical Considerations. *Journal of Gay & Lesbian Issues in Education*, 3(2-3):35–56, Oct. 2008.

[4] N. DiFonzo, J. Suls, J. W. Beckstead, M. J. Bourgeois, C. M. Homan, S. Brougher, A. J. Younge, and N. Terpstra-Schwab. Network structure moderates intergroup differentiation of stereotyped rumors. *Social Cognition*, 32(5):409–448, 2014.

[5] A. P. Haas, M. Eliason, V. M. Mays, R. M. Mathy, S. D. Cochran, A. R. D'Augelli, M. M. Silverman, P. W. Fisher, T. Hughes, M. Rosario, S. T. Russell, E. Malley, J. Reed, D. A. Litts, E. Haller, R. L. Sell, G. Remafedi, J. Bradford, A. L. Beautrais, G. K. Brown, G. M. Diamond, M. S. Friedman, R. Garofalo, M. S. Turner, A. Hollibaugh, and P. J. Clayton. Suicide and suicide risk in lesbian, gay, bisexual, and transgender populations: review and recommendations. *Journal of homosexuality*, 58(1):10–51, 2011.

[6] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[7] J. D. Hawkins, R. F. Catalano, and J. Y. Miller. Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: implications for substance abuse prevention. *Psychological bulletin*, 112(1):64, 1992.

[8] C. M. Homan, N. Lu, X. Tu, M. C. Lytle, and V. Silenzio. Social structure and depression in trevorspace. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 615–625. ACM, 2014.

[9] B. Latané and T. L'Herrou. Spatial clustering in the conformity game: Dynamic social impact in electronic groups. *Journal of Personality and Social Psychology*, 70(6):1218, 1996.

[10] J. S. McDaniel, D. Purcell, and A. R. D'Augelli. The relationship between sexual orientation and risk for suicide: Research findings and future directions for research and prevention. *Suicide and Life-Threatening Behavior*, 31(s1):84–105, 2001.

[11] B. Mustanski. Ethical and Regulatory Issues with Conducting Sexuality Research with LGBT Adolescents: A Call to Action for a Scientifically Informed Approach. *Archives of sexual behavior*, 40(4):673–686, Apr. 2011.

[12] C. Peersman, W. Daelemans, and L. Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2011.

[13] M. Pennacchiotti and A. Popescu. Democrats, republicans and starbucks afficionados: user

classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 430–438. ACM, 2011.

[14] M. Pennacchiotti and A. Popescu. A machine learning approach to twitter user classification. *ICWSM*, 11:281–288, 2011.

[15] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2010.

[16] R. L. Sell and V. Silenzio. Lesbian, gay, bisexual, and transgender public health research. In *The Handbook of Lesbian, Gay, Bisexual, and Transgender Public Health*, pages 33–56. The Handbook of Lesbian, Gay, Bisexual, And Transgender Public Health: A Practitioner's Guide to Service, New York, 2006.

[17] S. Stephens-Davidowitz. How Many American Men Are Gay? page SR5, Dec. 2013.

[18] K. Ueno, E. R. Wright, M. D. Gayman, and J. M. McCabe. Segregation in Gay, Lesbian and Bisexual Youth's Personal Networks: Testing Structural Constraint, Choice Homophily and Compartmentalization Hypotheses. *Social Forces*, 90(3):971–991, Mar. 2012.

[19] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

[20] X. Zhu, Z. Ghahramani, J. Lafferty, et al. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, volume 3, pages 912–919, 2003.