2015 IEEE 16th International Conference on Information Reuse and Integration

Gender Prediction in Random Chat Networks Using Topological Network Structures and Masked Content

Michael Crawford and Xingquan Zhu
Dept. of Computer & Electrical Engineering and Computer Science
Florida Atlantic University
Boca Raton, Florida USA 33431
michaelcrawf2014@fau.edu, xzhu3@fau.edu

***Abstract*—Social media is becoming a critical avenue for businesses today to target new customers and create brand loyalty. In order to target users effectively, companies need to know basic information about their users. However, in many cases, user profiles are either incomplete or completely wrong, and one of the most critical pieces of private information is gender. In this paper we examine the case of gender prediction in random chat networks using masked content and topological network structures. Random chat networks (*e.g.*, Chatous.com) are significantly different from most existing social networks, because users do not get to see who they are going to talk to before they engage in a conversation, and the system itself brings users into chats together. Due to the network's random nature, users have very little information about peers in the network. Additionally, privacy is an ever growing concern in today's society, thus data analytic tools often need to work with data which has been masked to prevent a possible breach of confidential information. In the paper, we first analyze some fundamental characteristics of random chat networks when broken down by gender. Then we propose an approach for gender prediction using masked words as features and show that gender prediction performance can be boosted by incorporating network topology statistics. Finally, we will examine network statistics which are most useful for gender prediction.**

***Keywords*—Gender Prediction; Masked Content; Topological Structure; Social Networks; Random Chat Networks; Random Forest; Data Mining; Web Intelligence**

I. Introduction

Social media is being used by more than one-seventh of the world's population [1] and is a new and powerful medium for businesses to target customers and create brand loyalty [2][3]. In social network environments, user profiles are often incomplete or inaccurate[4], raising the need to predict missing information or identify potentially falsified information[5]. One of the most important attributes of a person's profile is the gender, and as such, there have been several studies on predicting gender in social networks such as Facebook, Twitter, LinkedIn, YouTube, MySpace, Fotolog and NetLog [6][7][8][1].

In order to predict users' genders, one of the most important tasks is to find effective features for gender characterization. Existing methods are usually focused on using simple n-grams or Linguistic Inquiry and Word Count (LIWC), including preconceived notions of groups associated with words and phrases, as features. For short messages, such as SMS, chats, or tweets, solutions also exist to explore domain specific features, such as using abbreviations and emoticons [like ;) or :( ], which are unique for these types of media versus traditional written media [1].

A few studies have also focused on improving Twitter gender prediction by augmenting the standard LIWC and n-gram features with various derived communication behavior statistics, such as follower-following ratio, follower frequency, following frequency, response frequency, retweet frequency and tweet frequency. Interestingly, the study observed little to no additional performance gain by using these features [9].

Assuming suitable features are collected, many learning algorithms can be directly used for gender prediction. Examples include (but are not limited to) Support Vector Machines, Naive Bayes, Bayesian Logistical Regression and Decision Trees. In one particular study, a simple heuristic model based upon the user's username was used [4]. In this study, we will use Random Forest [10] to predict gender by using masked word content (between two users who engaged the conversation) and a conceived network based on previous chats with some statistics derived from the network.

Random chat networks[1] are unique social networks, where users do not actually know each other prior to engaging in a conversation. Users are placed into a chat together either randomly or based upon their common interests. Because of this random nature, users have very little information about the person they are speaking with or their peers in general. This unique random chat setting raises many interesting questions, such as what are the underlying network characteristics, in comparison to general social networks? Do males *vs.* females have different network signatures in such a random world? Can we predict user gender information from such random chat networks? In this paper, we intend to bring answers to these questions.

In our study, a set of 9 million chat logs from Chatous.com are used as our test bed. The data consists of a giant component of over 300,000 users where there is a link between any two users if they have chatted. For each conversation, the chat content is masked as a vector recording words users had spoken at any time while chatting. By examining the giant component of the network, various network statistics such as degree, betweenness centrality, clustering coefficient and page rank can be calculated. We will analyze these features to show that these statistics aid tremendously in predicting gender, compared to using masked word counts alone. Then, each of the derived

[1]An example of random chat network is Chatous.com

978-1-4673-6656-4/15 $31.00 © 2015 IEEE
DOI 10.1109/IRI.2015.35

features and combinations of these features will be evaluated to determine their relative importance. With this particular dataset, all words in chat sessions are masked (as tokens) to protect user privacy. So the standard LIWC features and part-of-speech tagging cannot be applied, making the system blind to what the words actually are. This is, in fact, important to realize, because due to privacy concerns, the raw content of messages are often unavailable to data analytics models. This is also different from privacy preserving data mining where data analytics are still able to work on anonymized data [11] in unmasked forms. To the best of our knowledge, this is the first study of gender prediction for social networks with masked content information, and networks where users have little to no information about who they are speaking with.

II. Related Work

Gender is a very useful, yet private, piece of information that can help infer many important clues, such as customer shopping interests and user behaviors. As a result, there have been many studies on predicting gender from different perspectives. Schwartz *et al.* investigated the differences between predicting personality, gender and age using Facebook status updates [1]. Specifically, they compared using an open-vocabulary approach (no preconceived notion of class for words or phrases) versus a closed-vocabulary approach (words and phrases are preclassified into groups), and suggested that open vocabulary approach can work better when used for prediction, but can be further improved by combining the two together.

Peersman *et al.*'s study, "Predicting Age and Gender in Online Social Networks", also dealt with predicting age and gender based upon words, but was limited to a Belgian website using n-grams varying from 1 to 3 [5] . For age, instead of trying to predict an exact age they are only trying to predict young versus old, *i.e.* a binary classification task. The purpose of the study was to try and identify pedophiles. However, as the author notes, trends for the general population may not generalize to the specific patterns of a pedophile and they would need to add specific test cases for them to see if they can spot patterns in their language.

Bamman *et al.* proposed to classify gender based upon an open dictionary approach [12]. The study begins with classifying gender based solely upon word occurrences. Next, they manually grouped the words into categories and tried to find clusters of users that used certain categories of words. An interesting observation was that certain predominately male or female groups often displayed stylistic differences from the general trends for male or female. For example, females use words like "lmao" at a higher rate than males. Meanwhile, they also found a group of males which used the term at a much higher rate than is normal for males and a corresponding group of mainly females, which used the term at a much lower rate. An opportunity for further research may be to train individual models based upon these groups and then use these new models for gender prediction.

In addition, Bamman *et al.* also investigated homophilous groups of friends, by using the percentage of male and female friends as an additional input attribute. They found that for the most part, males tend to have more male friends, and vice versa for females. They also noticed that females and males, with more female and male friends respectively, tend to have more textual attributes in their messages which are female or male. That is to say, a female with more female friends tends to have a higher occurrence of linguistic markers which signify them being female, and conversely, a male with more male friends tends to have a higher occurrence of markers signifying them as being male. Further research could be performed to test how training models based upon the distribution of female versus male friends affects the classification of gender in general.

Smith attempted to predict gender based upon a user's username by comparing it with known names for gender [4], and achieved an accuracy of 55% when using this method with both male and female names. However, when only using female names, along with the words "girl" and "love", he was able to obtain an accuracy of 64.7%. Of particular note here is that this indicates that females use their real name more often than males in their username, and while 64.7% is low, it is using an extremely limited amount of information, namely just the username. It would have been interesting to see what percentage of users which had a name that matched were correctly predicted since they randomly predicted users as female at a rate of 70% if they were not matched. Also this would have given some indication of the percentage of males using a female name as their user name.

III. Benchmark Data

The particular dataset used in this study is a set of over 9 million chats from over 300,000 users in Chatous.com (a typical random chat social network). Random chat networks are a unique type of social network, in that users do not actually know each other before talking, but are simply brought into chats randomly or based upon users' common interests.

From each of the chat logs, the following items are collected:

- First User Profile ID – The profile id of the first user of the chat which can be matched against the profile file.
- Second User Profile ID – The profile id of the second user of the chat which can be matched against the profile file.
- First User Word Vector – A list of words spoken by the first user. Words are encoded (*i.e.* masked) so that the actual words used in the conversation are unknown, preventing the employment of common natural language processing tools such as stemming or part-of-speech.
- Second User Word Vector – List of the words spoken by the second user, which are also masked.

The dataset also contains a profile for each user. From the profiles we extracted the following items:

- Profile ID – The ID of the profile which the chats can be cross referenced with.
- Age – The age of the user. Users with unknown age are marked as 0.

- Gender – The gender of the user. Users with unknown gender information are removed so that we know the genuine gender information of every user in the dataset since our gender prediction is based on the validation of the predicted gender value *vs.* the user provided gender value. The final counts of males and females is shown in Table I.

Table I: The number of male and female users in the preprocessed dataset

Gender	Count
Female	121461
Male	199392

IV. Algorithm Design

The general design of our framework is as follows:

1) Two files are created to extract items detailed in Section III. The files are passed into a set of python scripts to create the network. Each node of the network denotes a user and two users are connected through an edge if there is a conversation between them.
2) Once the network is built, the giant component of the network is extracted as detailed in Section V.
3) After the giant component of the network is collected, a set of python scripts were used in conjunction with the SNAP graph framework[13] to compute the network features as described in Section VII.
4) A dataset file is created using Python which includes network features of each node along with the list of masked words which had been spoken by the user. Section VI details the construction of these lists.
5) The experiments are then run using the WEKA experimenter as detailed Section IX.

V. Network Creation

The network is created with the definition that each user is a node in the network, and an edge connects two users if a chat log exists between them. In order to construct the benchmark network, we use a python script to go through the chat log and output an intermediate file which simply contains a row for each edge of the network in the form of [user1;user2].

After the above process, the edge list is passed into a second script which creates a TUNGraph class (undirected network) using the SNAP framework [13]. After that, the giant component of the network, which contains 331,572 nodes, is extracted from the network. This giant component is then saved into a file for further processing. The full list of components is detailed in Table II.

VI. Masked Word Vectors

In the chat log, a string records all words which had been spoken by any given user. Note that words are encoded/masked like `word:xxxxx` where `xxxxx` is simply a number so we do not know the actual words used in the conversation, leaving no part-of-speech or LIWC features available for analytics. This feature ends up simply being converted to a bag-of-masked-word vector for classification where the feature is 1 if a particular masked word is used in a conversation, or 0 otherwise.

Table II: The sizes and counts of the connected components in the generated network

Size	Count
331,572	1
11	1
4	8
3	20
2	606

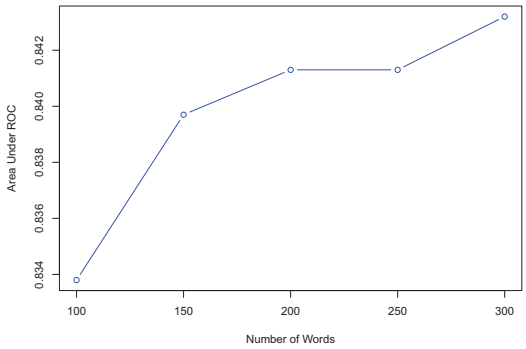

Figure 1: The Area Under ROC as the number of words in the word vector feature is increased

In order to decide the number of words to be included in the word vector, a short test is run using Random Forest [10] and comparing the Area Under the ROC curve versus the number of words included in this word vector. The results are shown in Figure 1, which demonstrates a drastic increase from 100 to 150 words and another slight bump from 150 to 200. There is virtually no difference between 200 to 250 while only a slight bump from 250 to 300. Therefore, we use 200 words as a nice middle ground to keep training time and classifier complexity down.

VII. Network Features

Several network features are computed from the entire network. We note that the test data is used in creation of the network features, however, this is inline with our particular use case of inferring gender in an existing network. Further, gender is not used in any way during the creation of the network statistics. We briefly outline these features in the following subsections:

A. Betweenness Centrality

The betweenness centrality score is defined by (1) where $P_i(k, j)$ is the number of shortest paths passing through node i, and $P(k, j)$ is the total number of shortest paths between k and j. For any pair of nodes, if none of their shortest paths pass through i, the node pair will bring 0 contribution towards the betweenness centrality score of node i. Due to the size of this particular network, the betweenness centrality was computed as an approximation using 10% of the nodes in the network.

The algorithm used in our experiments is implemented in the SNAP framework and is based upon Brandes and Pich's work "Centrality Estimation in Large Networks" [14].

$$C_B(i) = \sum_{k,j} P_i(k,j)/P(k,j) \quad (1)$$

B. Node Degree

The node degree is simply the number of edges that exist in the network from a particular node to any other. In our case, this means it is simply the number of other users with whom a user has chatted.

C. Clustering Coefficient

The clustering coefficient is the fraction of a node's neighbors which are connected and is given by (2).

$$C(v_i) = \frac{\text{Pairs of Neighbors of } v_i \text{ That Are Connected}}{\text{Number of Pairs of Neighbors of } v_i} \quad (2)$$

D. PageRank

PageRank score is a measure of a page importance. In order to calculate the importance score of a page x, the intuition of the PageRank is to check the importance scores of the pages which point to x, and then linearly combine their importance scores to form the important score for x.

A simple version of PageRank is formally defined by $R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$. In our particular case, we include a dampening factor d which was set to 0.85 and is formally defined in (3) where N_v is the number of links from page v [15]. While PageRank is designed for a directed network it can be easily adapted to an undirected network by simply having a directed link in both directions. The final score is computed using the power iteration method with maximum number of iterations being set to 100.

$$R(u) = \frac{1-d}{N} + d \sum_{v \in B_u} \frac{R(v)}{N_v} \quad (3)$$

VIII. Random Forest

Random Forest (RF) is an ensemble learner in which many unpruned decision trees are built [10]. Bagging is used to construct the trees in that a random subset (with replacement) of features and a random subset of data are selected to build each tree. While building the trees, a random subset of features are considered at each decision node. After all trees are built, classification takes place by evaluating the instance with respect to all trees and the decision is the one agreed by the majority of the trees (*i.e.* a majority voting approach).

Several parameters need to be set when using RF, including the number of trees to be built and the number of features to be considered at each node. Following previous empirical studies [16], we use $\log_2 M + 1$ for the number of features to be considered in each node.

To determine the optimal number of trees, an experiment was conducted with our particular dataset and the results are shown in Figure 2. When varying the number of trees from 50 to 300, RF shows significantly performance gain, but the amount of time needed for training increases dramatically. So we split the difference and use 200 trees which gives relatively good performance while still limiting the amount of training time to a feasible number.

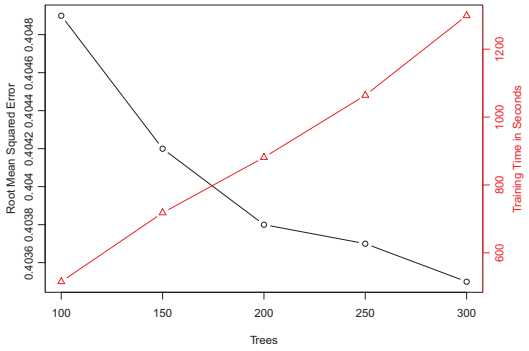

Figure 2: The Root Mean Squared Error as the number of Trees in Random Forest is increased along with the amount of time needed to train

IX. Experimental Settings

The main experiments conducted in this study are to validate whether we can accurately predict gender in random chat networks. If so, what are the important features tied to the prediction. Accordingly, we compare the experimental results by using different network features for gender prediction. The RF classifier used in this study was implemented in WEKA [17] and all of the default parameters are used except for NumTrees (200) and NumFeatures ($\log_2 M + 1$) as described previously. In order to construct the masked word vector, the meta.FilteredClassifier was used in conjunction with the StringToWordVector filter. For each iteration, the StringToWordVector filter is run on the training dataset to select the top 200 word features, with the same set of words being used for test.

All experiments are carried out using 10 times five-fold validation, with 20% of the data being used for training (64,170) and 80% for testing (256,682). In each case, the instances used for training are selected at random.

X. Performance Measures

Multiple performance measures are used in this study and are briefly described in the following sections.

A. Accuracy (ACC)

Accuracy is the number of correctly classified instances divided by the number of all classified instances, as defined in (4).

$$accuracy = \frac{TP + RN}{TP + FP + FN + TN} \quad (4)$$

B. Receiver Operating Characteristic (ROC) Curve

The ROC curve is a graph with the number of true positives on the Y axis versus false negatives on the X axis. The TP and FP rates are inversely correlated, and their values are normally collected by varying the parameter or threshold values. By determining the area under the ROC curve (AUC), one can effectively estimate the true predictive capability of a given classifier.

C. Root Mean Squared Error (RMSE)

The RMSE is the square root of the sum of all of the squared errors divided by the number of instances and is defined formally by (5).

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}} \tag{5}$$

XI. Results

In the following sections, we report some major findings made in our study. We will first investigate whether and how network features can be used to assist gender prediction. To this end, we will compare classifier performance with and without network features and validate how different combinations of network features are able to boost the performance of gender prediction in random chat networks.

A. Random Chat Network Characteristics

1) Degree: Figure 3 shows the counts of the different degrees broken down by gender. We can observe a few interesting trends here:

1) Overall, the degrees are similar to a scale-free network and the degree distributions are reasonably close to a general power law distribution.
2) There are more females with a very small node degree. However, as node degree increases the males overtake the females and the gap seems to widen as the degree increases.
3) A minor concern about the above conclusion is that the number of males (199392) is actually greater than the number of females (121461) as was shown in Table I. So we randomly sampled 121461 males from the population and compared them with the females as shown in Figure 4. We can see that there are still more females than males at the lower degrees and there is a switch at just over 10 and then the males tend to dominate. Note that this under-sampling is only used for investigation of the various network features, and we use the original dataset distribution when performing classification.

2) PageRank: Figure 5 reports the distribution of page ranks with the sampled data, which shows that the PageRank distributions also follow the same type of scale-free distributions as the degree. At the lower PageRank scores, there are more females, however, as the page rank score increases the males overtake the females and the gap continues to grow as PageRank increases.

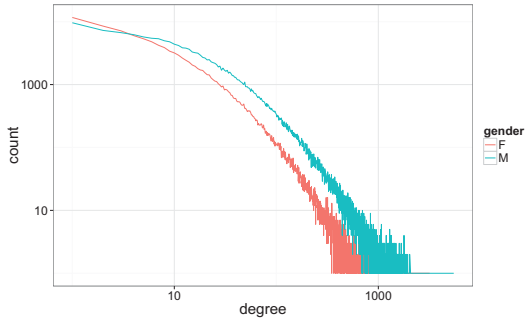

Figure 3: Degree counts broken down by gender on a Log-Log scale

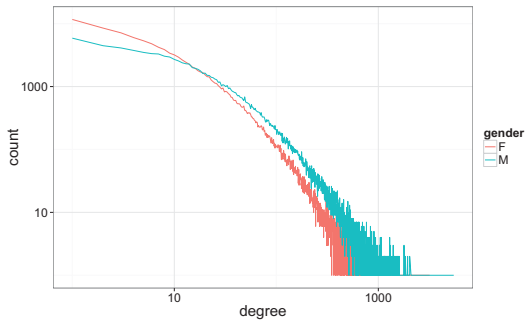

Figure 4: Degree counts broken down by gender on a Log-Log scale while undersampling the males to get a 50/50 distribution

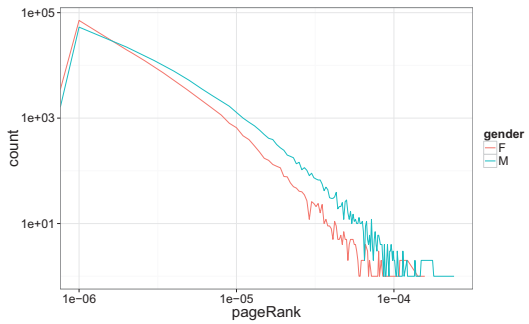

Figure 5: PageRank counts broken down by gender on a Log-Log scale while undersampling the males to get a 50/50 distribution

3) Clustering Coefficient: Figure 6 reports the clustering coefficient density plot, which shows that there is a noticeably higher density of females that have a clustering coefficient near zero while the highest density of males is around 0.2. Also of note here is that the males, in general, seem to be spread out more than the females; however, this does not seem to be as informative of a statistic as some of the others which will be evaluated more thoroughly in Section XI-C.

4) Betweenness Centrality: Figure 7 reports the density plot of the betweenness centrality scores, which shows that for the most part, both genders exhibit a relatively normal dis-

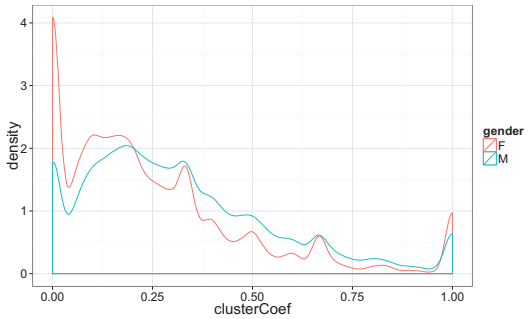

Figure 6: Density plot of Clustering Coefficient broken down by gender

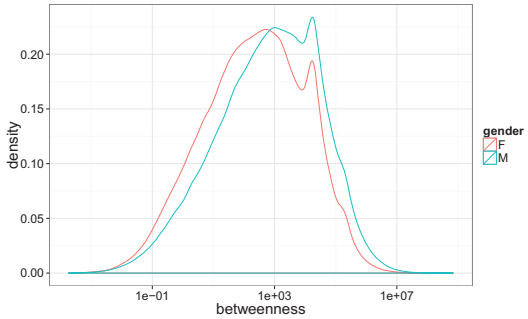

Figure 7: Density plot of Betweenness Centrality broken down by gender

Table III: Results of using network features versus not

Dataset	AUC	STDEV	ACC	STDEV
Age + Words	0.7988	0.00100	0.7509	0.00068
Age + Words + Network Features	0.8414	0.00086	0.7809	0.00103

ANOVA	DF	SS	MS	F	p-value
AUC	1	0.0091	0.0091	10422	<2e-16
ACC	1	0.0044	0.0044	5893	<2e-16

tribution. However, males do seem to have a slightly negative skew. As with clustering coefficient this does not seem to be as informative of a statistic as some of the others and will be evaluated more thoroughly in Section XI-C.

B. Gender Prediction Results with Network Features

The first, and most obvious test is to simply compare the performance of the RF classifier when using all computed network features versus the word vector. The results in Table III clearly show that network features are indeed helpful for gender classification. When performing an ANOVA analysis on the resulting AUC and ACC scores, the p-values are less than 2^{-16} indicating that the results are statistically significant.

C. Impact of Different Network Features

In this section, we compare different network features to determine which ones are the most informative for gender

Table IV: Comparison of the usefulness of each network feature

Network Feature	AUC	STDEV	ACC	STDEV
None	*0.7988*	0.00100	0.7509	0.00068
Betweenness	0.8150	0.00101	0.7675	0.00095
PageRank	0.8303	0.00101	0.7773	0.00071
Clustering Coef	0.8135	0.00109	0.7577	0.00065
Degree	**0.8395**	0.00096	0.7822	0.00039

ANOVA	DF	SS	MS	F	p-value
AUC	4	0.010024	0.0025060	2423	<2e-16
ACC	4	0.006821	0.0017054	3497	<2e-16

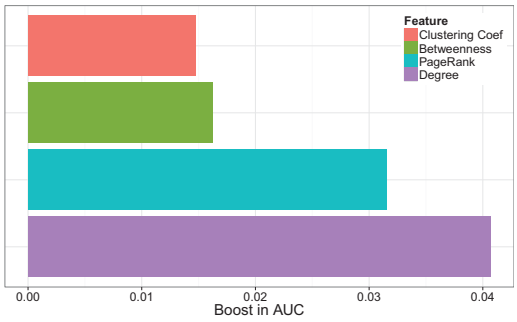

Figure 8: The boost in AUC score provided by each of the features individually

prediction. Table IV reports the results where each row corresponds to the classifier performance by using a specific type of network feature (and None means that no network features are involved). As can be seen in the table, degree is considered the most useful feature and the results of using different features are statistically significant. The boost in AUC score seen by including each of the features individually is shown in Figure 8.

In order to further compare each of the features, a pair Tukey's Honestly Significant Difference tests was run with an alpha of 0.05 and 0.01 and are shown in Table V. When looking at the AUC scores with an alpha of 0.5, Degree is significantly better than all others followed by PageRank, Betweenness, Clustering Coefficient and finally None. Overall, our observations suggest that each network feature provides significantly better performance than using no network features at all. When moving to an alpha of 0.1, the difference between Betweenness and Clustering Coefficient is not statistically significant. Looking at accuracy, the same order holds. Even with an alpha of 0.01, all features provide statistically significant different levels of accuracy.

D. Combinations of Network Features

Taking the analysis one step further, all combinations of network features are tested and the results are summarized in Table VI, which confirm that all combinations of network features perform better than no features at all (0000). Interestingly, using Degree, Clustering Coefficient and PageRank by themselves seems to perform better than also including Betweenness.

To further verify the significance of these differences, a Tukey's Honestly Significant Difference test was run and

Table V: Tukey's Honestly Significant Difference test of ACC and AUC across the pairs of features with alpha set to both 0.05 and 0.01

alpha = 0.05			alpha=0.01		
Feature	AUC	HSD	Feature	AUC	HSD
Degree	0.8395	A	Degree	0.8395	A
PageRank	0.8303	B	PageRank	0.8303	B
Betweenness	0.8150	C	Betweenness	0.8150	C
Clustering Coef	0.8135	D	Clustering Coef	0.8135	C
None	0.7988	E	None	0.7988	D
Feature	ACC	HSD	Feature	ACC	HSD
Degree	0.7822	A	Degree	0.7822	A
PageRank	0.7773	B	PageRank	0.7773	B
Betweenness	0.7675	C	Betweenness	0.7675	C
Clustering Coef	0.7577	D	Clustering Coef	0.7577	D
None	0.7509	E	None	0.7509	E

Table VI: AUC and ACC scores of all combinations of network features

Features	AUC	STDEV	ACC	STDEV
0000	*0.7988*	0.00100	*0.7509*	0.00068
0001	0.8150	0.00101	0.7675	0.00095
0010	0.8303	0.00101	0.7773	0.00071
0011	0.8220	0.00124	0.7725	0.00089
0100	0.8135	0.00109	0.7577	0.00065
0101	0.8322	0.00087	0.7777	0.00070
0110	0.8363	0.00116	0.7795	0.00096
0111	0.8374	0.00111	0.7812	0.00119
1000	0.8395	0.00096	0.7822	0.00039
1001	0.8355	0.00080	0.7800	0.00107
1010	0.8412	0.00114	0.7828	0.00090
1011	0.8352	0.00086	0.7777	0.00097
1100	0.8415	0.00097	0.7831	0.00104
1101	0.8417	0.00090	0.7831	0.00080
1110	**0.8434**	0.00100	**0.7835**	0.00120
1111	0.8414	0.00086	0.7809	0.00103

Features column is a bit encoded string where it is represented by: Degree, Clustering Coef, PageRank, Betweenness respectively.

Table VII: Tukey's Honestly Significant Difference test of AUC across the combinations of features with alpha set to both 0.05 and 0.01

Features	AUC	.05 HSD	.01 HSD
1110	0.8434	A	A
1101	0.8417	B	AB
1100	0.8415	B	B
1111	0.8414	B	B
1010	0.8412	B	BC
1000	0.8395	C	C
0111	0.8374	D	D
0110	0.8363	DE	DE
1001	0.8355	E	E
1011	0.8352	E	E
0101	0.8322	F	F
0010	0.8303	G	G
0011	0.8220	H	H
0001	0.8150	I	I
0100	0.8135	I	I
0000	0.7988	J	J

the results for AUC and ACC are shown in Tables VII and VIII. For AUC at the alpha = 0.05 level, the combination of Degree, Clustering Coefficient and PageRank by themselves is statistically better than all other combinations of features. However, when moving to alpha = 0.01, it is tied with swapping betweenness and clustering coefficient, which makes sense, since back in our previous test in Table V, Betweenness and Clustering Coefficient were not different in statistically significant way at the alpha = 0.01 level with AUC. When moving over to the accuracy scores we observed much less

Table VIII: Tukey's Honestly Significant Difference test of ACC across the combinations of features with alpha set to both 0.05 and 0.01

Features	ACC	.05 HSD	.01 HSD
1110	0.7835	A	A
1101	0.7831	A	A
1100	0.7831	A	A
1010	0.7828	A	AB
1000	0.7822	AB	ABC
0111	0.7812	BC	BCD
1111	0.7809	BC	CDE
1001	0.7800	CD	DE
0110	0.7795	D	E
0101	0.7777	E	F
1011	0.7777	E	F
0010	0.7773	E	F
0011	0.7725	F	G
0001	0.7675	G	H
0100	0.7577	H	I
0000	0.7509	I	J

variation at the top with 5 different features combinations all tying for the top spot.

XII. Conclusion

In this study, we investigated gender prediction in random chat networks by using network topology statistics. Random chat networks are quite different from most other well-studied social networks, since the system places users in chats rather than the users themselves, as is common in existing social networks. In addition, all words in our dataset are masked, removing contextual information, which is necessary in certain environments due to increasing privacy concerns. Our study found that by using network statistics, we are able to predict gender significantly better than using masked word vector features alone. Furthermore, our experiments show that in this particular network, Degree is the most useful feature for boosting gender prediction, followed closely by PageRank. While Betweenness Centrality and Clustering Coefficient are found to be less informative than others for gender prediction, they are still significantly better than using no network features at all.

Future work should entail comparing the network statistics of this particular network with other social networks to determine how the way in which users are placed into chat affects the overall network statistics. Also, the effect of including additional features which are available in some cases (*e.g.*, location, whether a person wants their location to be known, length of chats) should also be studied.

XIII. Acknowledgement

This research is partially supported by National Science Foundation under grant No. CNS-1427536. We sincerely thank Kevin Guo, CEO of Chatous, for providing the dataset for experiments and validations.

References

[1] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, "Personality, gender, and age in the language of social media: The open-vocabulary approach", *PLOS ONE*, vol. 8, no. 9, T. Preis, Ed., e73791, Sep. 25, 2013, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0073791. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0073791.

[2] M. Laroche, M. R. Habibi, and M.-O. Richard, "To be or not to be in social media: How brand loyalty is affected by social media?", *International Journal of Information Management*, vol. 33, no. 1, pp. 76–82, Feb. 2013, ISSN: 0268-4012. DOI: 10.1016/j.ijinfomgt.2012.07.003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0268401212000916.

[3] R. W. Naylor, C. P. Lamberton, and P. M. West, "Beyond the "like" button: The impact of mere virtual presence on brand evaluations and purchase intentions in social media settings", *Journal of Marketing*, vol. 76, no. 6, pp. 105–120, Nov. 2012, ISSN: 00222429.

[4] J. Smith, "Gender prediction in social media", *ARXIV preprint arXiv:1407.2147*, 2014. [Online]. Available: http://arxiv.org/abs/1407.2147.

[5] 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*, ACM, 2011, pp. 37–44. [Online]. Available: http://dl.acm.org/citation.cfm?id=2065035.

[6] N. Cheng, R. Chandramouli, and K. P. Subbalakshmi, "Author gender identification from text", *Digital Investigation*, vol. 8, no. 1, pp. 78–88, Jul. 2011, ISSN: 1742-2876. DOI: 10.1016/j.diin.2011.04.002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1742287611000247.

[7] K. Filippova, "User demographics and language in an implicit social network", in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ser. EMNLP-CoNLL '12, Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 1478–1488. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390948.2391117.

[8] A. Kokkos and T. Tzouramanis, "A robust gender inference model for online social networks and its application to LinkedIn and twitter", *First Monday*, vol. 19, no. 9, Aug. 29, 2014, ISSN: 13960466. [Online]. Available: http://firstmonday.org/ojs/index.php/fm/article/view/5216.

[9] 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*, ser. SMUC '10, New York, NY, USA: ACM, 2010, pp. 37–44, ISBN: 978-1-4503-0386-6. DOI: 10.1145/1871985.1871993.

[10] L. Breiman, "Random forests", *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://link.springer.com/article/10.1023/A:1010933404324 (visited on 10/29/2014).

[11] R. Agrawal and R. Srikant, "Privacy-preserving data mining", in *ACM Sigmod Record*, ACM, vol. 29, 2000, pp. 439–450.

[12] D. Bamman, J. Eisenstein, and T. Schnoebelen, "Gender in twitter: Styles, stances", andsocial networks. Technical Report 1210.4567, arXiv, October, 2012. [Online]. Available: http://www.cc.gatech.edu/~jeisenst/papers/GenderInTwitter923.pdf.

[13] J. Leskovec. (2014). Snap.py - SNAP for python, [Online]. Available: http://snap.stanford.edu/snappy/index.html (visited on 12/09/2014).

[14] U. Brandes and C. Pich, "Centrality estimation in large networks", *International Journal of Bifurcation & Chaos in Applied Sciences & Engineering*, vol. 17, no. 7, pp. 2303–2318, Jul. 2007, ISSN: 02181274.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web.", 1999. [Online]. Available: http://ilpubs.stanford.edu:8090/422.

[16] T. M. Khoshgoftaar, M. Golawala, and J. V. Hulse, "An empirical study of learning from imbalanced data using random forest", IEEE, Oct. 2007, pp. 310–317, ISBN: 0-7695-3015-X, 978-0-7695-3015-4. DOI: 10.1109/ICTAI.2007.46. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4410397 (visited on 10/29/2014).

[17] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*, 3 edition. Burlington, MA: Morgan Kaufmann, Jan. 20, 2011, 664 pp., ISBN: 9780123748560.