Age and Geographic Inferences of the LiveJournal Social Network

Ian MacKinnon
imackinn@uwaterloo.ca

Robert Warren
rhwarren@uwaterloo.ca

David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario

June, 2006

Abstract

Online social networks are often a by-product of blogging and other online media sites on the Internet. Services such as LiveJournal allow their users to specify who their "friends" are, and thus a social network is formed. This paper will explore the relationship between users with the intent of being able to make a prediction of a users age and country of residence based on the information given by their friends on this social network.

1 Introduction

Ever since the classic experiment of Milgram (Milgram, 1967; Travers & Milgram, 1969), people have taken an interest in people and the relationships that bind them. The field of social network analysis is growing with a number of interesting research problems and applications being studied. These have included the destabilization of terrorist networks (Carley et al., 2003) and the

study of information flow in organizations (Coleman et al., 1966; Rapoport, 1953).

We review here the preliminary results of our analysis of a partial LiveJournal data set. One can imagine the social network formed by LiveJournal as a graph, with each user being a vertex, and an edge existing between two vertices as an individual declares another user to be his friend. The findings here represent the first stages of a larger project to analyze the people and relationships that bind them online.

It is our intention to show a strong linking between the global location of users in LiveJournal with countries of their friends. The obvious application of knowing how strong this bond is would be to infer where a user is located based on information we know about their friends.

We also intend to look at the relationship between the age of a user and the age of their friends. LiveJournal stats[1] tell us there is a huge bias towards users being in their teens or early twenties. We first begin with general observations about the data and the relationships contained within it and then present our work on location analysis.

2 Background

Initial user discovery was accomplished by polling the LiveJournal "Latest Posts" [2] feed for the first week of September 2005. Of users discovered, only users identifying themselves from the top 8 countries in Livejournal [3] will be analyzed in this paper. Table 1 shows the number of users from each of the top 8 countries as given by LiveJournal at the end of user discovery.

Using these initial users as a start, a breadth-first search was performed by crawling the users information page on LiveJournal, identifying the users 'friends', queueing the friend to be crawled if (s)he did not already exist in the system, adding an edge from the first user to the friend, and continuing. For each user crawled, any demographic data volunteered by the user was recorded.This process was performed at small intervals over the course of September to December 2005.

[1] `http://www.livejournal.com/stats.bml`

[2] `http://www.livejournal.com/stats/latest-rss.bml`

[3] `http://www.livejournal.com/stats.bml`

Table 1: Top 8 countries of origin reported by users on LiveJournal.

Country	Count
United States	2990918
Russian Federation	252258
Canada	233839
United Kingdom	191650
Australia	89729
Philippines	31230
Germany	29224
Ukraine	28478

2.1 Data Collection

In total, information from 4,138,834 LiveJournal users were collected, 2,317,517 bloggers from our sample had reported to be of the top 8 nations in LiveJournal listed in Table 1.

As identified by (Lin & Halavais, 2004), there are 2 limiting factors when dealing with self-reported data: Many users do not report their location or age or report erroneous or false data.

Since users may lie about their age or location, there are a number of implausible or unlikely situations in the data. These range from the 112-year-old blogger to the school girl with a network of Michigan friends listing Afghanistan as a home. We will consider users who lie about their age or location to be "noise" in our data set.

We hope that with a better understanding of the relationships within the data, we can best identify users for whom their reported information is false. One element that we initially wished to review was the associativity of relationships: were friendships mutual? Or were people randomly linking to each other as friends?

2.2 Are Relationships Associative?

In the LiveJournal social network, as user u can declare user v a friend, without v reciprocating the declaration. In some online social networks such as Facebook, friend status is symmetric in that u is a friend of v if and only

if v is a friend of u. We will now look into how many friend relationships between users are reciprocated in our data set.

Because our sample of the LiveJournal database was limited, we eliminated any relationship from the dataset that linked to an undiscovered person within the universe. This was done to ensure that there existed the possibility that any recorded relationship be reciprocated.

Adjusting for these unknown persons, we recorded 36119462 (70%) symmetric relationships and 16082909 (30%) asymmetric relationships, or 52,202,371 known relationships linking 4,752,296 bloggers.

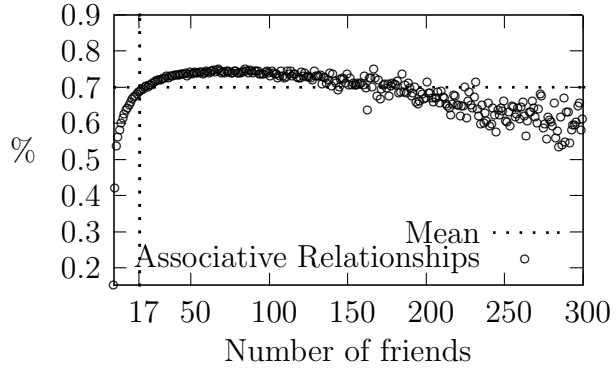

Figure 1: Percentage of a person's relationship that are associative against size of friend list.

Figure 1 plots the ratio between the number of symmetric relationships to asymmetric relationships and clusters the results by the size of the declared circle of friends. On average, a user has 17 friends on LiveJournal, 70% of which consider them friends in return. Our results are comparable to (Kumar et al., 2004), which reported 80% of relationships being associative.

What we found interesting was the low number of symmetric relationships for users with a small circle of friends (< 17). Initially, we thought that this was an artifact of a bad sample or that it was due to user logging onto the system once and never following up to new friendships.

We attempted to account for both these situations by observing the distribution of the number of friends per user and attempting to remove users that had posted only once or less to their blog. In both cases, we found that the distribution and the removal of stale users did not change this low number of associative relationships.

The greater variance in the ratio for larger number of friends (> 200) is likely to be due to the mechanical difficulties in managing a large number of friends within the blogging applications and/or due to "popularity" effects around certain individuals. We have already observed this last phenomena in networks of email cryptography keys, where certain individuals have a disproportionate number of in-links from other people claiming to be their friends (Warren, 2005). We expected this phenomena to occur with much smaller circle of friends.

2.3 Is Age a Factor in Relationships?

Intuitively, it seems reasonable that on average people should have a circle of friends that is roughly of the same age as them. Within our dataset, the average age of a LiveJournal user is 23 years also.[4] Figure 2 plots the distribution of the different age groups within the dataset and expresses that the bulk of the population is ages between 15 and 35 years old.

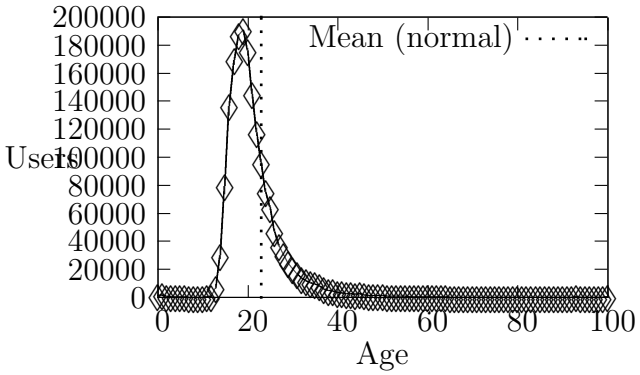

Figure 2: Histogram of self-reported age in the dataset.

A question that we wished to answer was whether it was possible to relate a person's age with the age of the people that he associates with. If there exists an age-group relationship, could we exploit it to rebuild missing data or to identify erroneous age data within the dataset?

[4]We assume here a normal distribution for comparison with a users social network. This is clearly wrong according to the distribution of Figure 2, however the normal mean is an intuitive measure to compare against for small age groups.

We first gathered age data for all of the users within the dataset (effective Dec. 31, 2005) and calculated the average age of the friends for each individual user. We also gathered the standard deviation for each group of friends and clustered the information according to users age.

There exists a direct relationship between both a person's age and the mean of the age of her friends. Large variations in the mean and standard deviations of users of age 60 and over can be explained by both a limited sample and a high likely hood that the age data is erroneous. The data concerning users less than 20 years old is cause for concern because of the high standard deviation and the localized drop in the mean age of the group of friends. We currently have two hypotheses that we will verify in future work: 1) that a number of younger LiveJournal users are declaring family members to be friends, which would account for age variations and 2) that there may exist a tendency for younger users to report their age as closer to the mean age.

We did attempt a simple slope classifier based on the normal age mean of the immediate social circle, divided by the user's age. To avoid any issues with erroneous data for this proof of concept, we chose all age classes that occurred more than 4,000 times within the dataset, dropped any record with missing information. We retained 1,022,400 users, aged from 17 to 39 years old, along with their social network.

We randomly split the data in two equal sets of 516,200 users and calculated the mean age of their social network. Using the first set, we calculated the average slope of the user age versus the mean social network age, which is 0.992. We then used the second set to benchmark the precision of the classifier at different prediction interval.

Table 2 represents the different precisions obtained within certain confidence intervals. For convenience, we represents the interval in time periods. Through experimentation with linear regressions and other classifications methods, we have concluded that there does exist a relationship between the age of a person and their peer group. However, the results in Table 2 are typical in that a certain number of user cannot be easily classified. Based on an inspection of the data, we believe that this is because there are several classes of users that require different classifiers. Hence, a child with a online blog may link to much older family members and a classifier capable of dealing with exceptions will be needed to handle these cases.

We believe that this may be an opportunity for us to both identify the class or type of blogger (e.g.: child, professional, parent). Furthermore, these

Table 2: We can predict the age of a core of users based on the mean age of their peer group.

Age Range(+/-)	Precision
4 months	0.18
6 months	0.29
8 months	0.39
1 year	0.49
1 1/2 years	0.62
2 years	0.71
2 1/2 years	0.76
3 years	0.80
3 1/2 years	0.83
4 years	0.86
5 years	0.98

features could be used in reverse to identify erroneous or fraudulent data with social networks.

3 Inferring Country Location

In this section, we review some of our initial results on inferring the home country of a blogger based on their social circle. Our intent is to create a lookup table of probabilities, where we can lookup what percentage of a users friends are from a country, X, and see what the probability is that a user is also from country X. For example, if we know that 30% of a user's friends are Canadian, what is the probability that the user is Canadian himself?

Of all the data collected, we will only be looking at those for whom we have location data for at least 10 friends. This is mostly to prevent cases where a user with a small number of friends causes great variance in the resulting probability ranges due to a single relationship.

Of the reduced sample size, there are a total of 5,831,566 "friend" edges with users in different countries. Of these, the United States accounts for 1,895,783 (32.5%) of all of these. Canada came in second with 630,478 (10.8%), followed by UK and Russia with 591,957 (10.1%) and 521,906 (8.9%)

respectively.

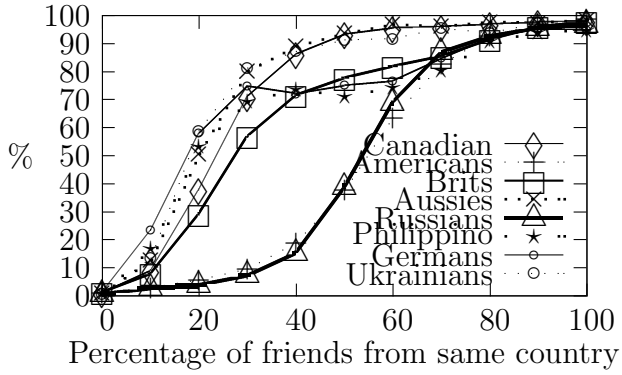

Figure 3: Probability of country of residence based on the dominant country of friends.

3.1 Sample

We are only looking at relationships between users who have identified themselves as being in the top 8 countries, and their friends who are also in one of these top 8 countries. Although this restricts our problem domain, the method can be extended to predict other nationalities. Hence, if a user had a number of friends in the top 8 countries but was not from one of the top 8 countries himself, they would not be included in this study. For simplicity, we group our results into ranges of 10%.

This does bring up an interesting problem that this field of research must deal with. Since obtaining a LiveJournal account is not mandatory, it represents only a certain subset of the world population. Furthermore, small world effects can un-intentionally skew results. Take for example the disproportionate Brazilian population within the Orkut social network site which dwarfs the United States. Hence the question is always whether we are dealing with a proper sample of the world, or if the dataset should be dealt with as a world in itself. In our work, we make the simplifying assumption that we are dealing with a world.

For simplicity, we cluster results in ranges of 10%, thus a user should have the possibility of being placed into any of these groupings. For example, if

a user only had 4 friends, it would be impossible for them to have friends in the 40-49% range.

3.2 Classification Table Creation

With a sub-graph of our sampled data fr Table 1 for which we develop probabilities for. It is because of cutting of users without geographic data and those who are not in the top 8 that we have the total same size given in Table 1 instead of the roughly 4 millions user records our discovery method had recovered.

Given our new subset of information, we iterate through all the users and for each country and divide the number of friends that user has from that country by their total number of friends.

We now have the percentage of each user's friends that are from each country. We then group these percentages (excluding 0%) into 10% ranges. Thus, we can establish a general trend in regards to the relationship between people with these percentage of friends from a country, and the probability they are also from that country.

To do this, we consider all the users in a range, and divide the number of users in this range who are from the country being considered by the total number of people. We thus have the probability that the user is in the country himself.

Repeating this for every user in every country of the top 8 yields the probability tables listed in Appendix Table 3-10.

3.3 Analysis of Classification Table

Most countries in the top 8 follow a roughly similar curve: there is a high degree of certainty about their nationality once half of their friends are from that country. Canada, UK, and Australia show this trend the most: there is an 80% chance of a user being from one of three countries if more than half of their friends are.

Interestingly, we see that Americans and Russians have a radically different curve from the rest. This indicates that many people who are not American or Russian, have a large number of American or Russian friends. For example, users with 50% Russian friends has less than a 20% chance of being Russian themselves.

One would expect this trend from the United States, seeing as how the number of American LiveJournal users dwarfs the number of users from any other country: statistically Americans should be linked to just about everyone. However, it is curious to see the same trend in Russia, but not Canada, which has a similar relative representation to the United States within the LiveJournal population.

3.4 Cross National Friendship Analysis

In order to get a better understanding of the reasons why the Russian and American curves are so radically different, we began looking at the friendship edges between the nations.

If we look at all friend edges that have a endpoint in Russia we find that there is a total of 2,573,840 of these edges. 2,051,934 of these are Russia to Russia edges (79.7%), 90,695 are Ukraine to Russia edges (3.5%), 66,539 are US to Russia edges (2.6%), and 42,788 are Israel to Russia edges (1.7%).

We compare this to American edges, for which there are a total of 21,695,176 edges where a "friend" is in the United States. 19,799,393 of these are US to US edges (91.3%), 482,965 are Canada to US edges (2.2%), 347,426 are UK to US edges (1.6%) and 162,077 are Australia to US edges (0.7%).

We see that the US and Russia both have the vast majority of their links within their own country. However, when we compare with the Canadian and Ukrainian edges we begin to see how this phenomena occurs.

Out of 1,469,750 edges that end in Canada, 839,272 are Canada to Canada edges (57.1%) and 482,583 are US to Canada edges (32.8%). For the 234,678 edges that end in the Ukraine, 117,982 are Ukraine to Ukraine (50.3%), and 77,098 are Russia to Ukraine (32.9%).

As we can see, the US/Canada relationship mirrors the Russia/Ukraine relationship in that the linkings between the two nations are very small in terms of the larger nation, but are quite significant for the smaller nation. Thus, there are many Canadians with a lot of American friends, and there are a lot of Ukrainians with many Russian friends. In both the US and Russia, the higher percentiles ranges have far more users than the the lower ones. Hence, the lower percentile ranges are diluted by the smaller friendly nations and we begin to see how the curve for the Russians and Americans can be shifted.

3.5 Application of Matrix to "Under 10 Friends"

From the initial sample set, any user with less than 10 friends for whom we have country data was not analyzed. This was meant to reduce variance in our results, but we wished to show that this decision would not have an adverse affect on our results. To do this, we applied our results from the classification table to this set of users who have less than 10 friends with known country of origin.

Specifically, for each user with less than 10 friends, we iterate through the classification tables for each country of the top 8, and saw what percentage of the users friends are from the country, lookup that country in the table and get the probability that user is from that country. We keep track of the probability the user is in each of the top 8 countries, and guess the country with the highest probability.

When we ran this procedure on the 1,024,944 users for whom we had country of origin data and less than 10 friends, we found that 143,899 were misclassified, or an accuracy of 86.0%. Hence, our simplifying assumptions were not overly affecting our end classifier.

3.6 Accuracy of Guessing Nationality

Finally, we attempted to determine the performance of the classifier under different information constraints, specifically the size of the immediate friends set. We grouped the users by their number of friends for whom we have country information and classified them.

This yielded Figure 4 where the average precision is plotted against the number of friend that the user has.

From this we see that the more friends a user has, the less accurate our ability to predict their country is. This result seems to indicate that as a user gets more friends, they get more and more from outside their home country. Hence, it becomes harder to predict their nationality as their friends are coming from a more diverse as they come into contact with more people.

4 Conclusions and future work

If the lookup probability model were performed on all countries represented in LiveJournal social network instead of just the top 8 countries, we could have a effective way to guess the location of a blogger based on their friends.

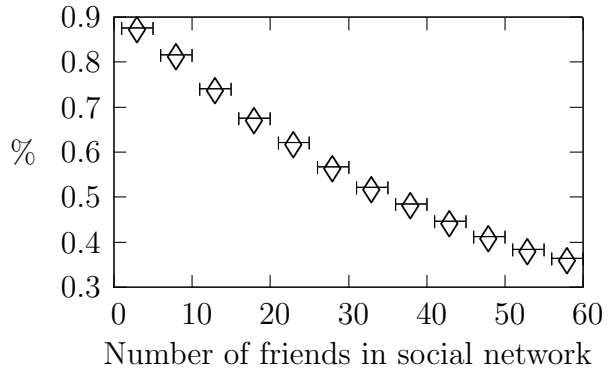

Figure 4: Country of origin naive predictor precision versus number of friends.

However, if other countries continue to follow trends, we may be able figure a "general" trend for other countries without having a specific lookup for each country. America and Russia have their own friend probability trends, but the others all follow fairly close to one another.

The model shown above would be more credible if it's conclusions could be verified on a similar social network such as Orkut or Xanga. Social networks such as Facebook should be avoided though, as Facebook is specific to a post-secondary institution, and was designed to create a social network within a University, rather than with people on the Internet at large because of bias towards those already residing in the same nation.

Hurst has shown how different blogging services can have radically different user bases (Hurst, 2005), so it would be interesting to see if international linkings followed a similar pattern on this different services and whether the distinct American and Russian trends are repeated.

LiveJournal has a fairly young user base, and if services such as Orkut were found to have a higher average age of user, we may see more variance in the location of users as adults could conceivably have more friends abroad they had early met in University or in the workplace.

We wish to extend the age and location models presented here to not only find falsely reported location and age information, but also correct it.For example, US/Canada and Russia/Ukraine linkings have been shown to be quite strong. Hence, if a person has many friends from the US, but does not claim to be in the US or Canada, this may indicate an error or dishonesty.

We also wish to integrate the "Link Prediction Problem" described by (Liben-Nowell & Kleinberg, 2003) to not only consider the information given, but also the relationship information that will occur in the future based on the current graph (Feld & Elmore, 1982).

5 Acknowledgments

The authors would like to acknowledge the help of Jonathan Fishbein in acquiring and pre-processing the data.

References

Carley, K. M., Reminga, J., & Borgatti, S. (2003). Destabilizing dynamic covert networks. *Proceedings of the 8th international Command and Control Research and Technology Symposium.*

Coleman, J. S., Katz, E., & Menzel, H. (1966). *Medical innovation: A diffusion study.* Bobbs-Merrill.

Feld, S. L., & Elmore, R. (1982). Patterns of sociometric choices: Transitivity reconsidered. *Social Psychology Quarterly*, *45*, 77–85.

Hurst, M. (2005). Gis and the blogosphere. *2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics.* Chiba, Japan.

Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2004). Structure and evolution of blogspace. *Commun. ACM*, *47*, 35–39.

Liben-Nowell, D., & Kleinberg, J. M. (2003). The link prediction problem for social networks. *Conference on Information and Knowledge Management (CIKM 2003)* (pp. 556–559). New Orleans, Louisiana, USA.

Lin, J., & Halavais, A. (2004). Mapping the blogosphere in america. *Workshop on the Weblogging Ecosystem at the 13th World Wide Web Conference (WWW2004).* New York City, USA.

Milgram, S. (1967). The small world problem. *Psycology Today*, 61–67.

Rapoport, A. (1953). Spread of information through a population with sociostructural bias. *Bulletin of Mathematical Biophysics*, 523–543.

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, *32*, 425–443.

Warren, R. H. (2005). The case for the dynamic analysis of social networks. *National Program on Complex Data Structures Workshop on Data Mining*. Toronto, Ontario, Canada: Fields Institute.

Appendix

Table 3: Percentage of American friends to probability of being American

Percent Range	American	Total	%
0-0.09	398	33114	1.2
0.1-0.19	443	15115	2.9
0.2-0.29	560	10006	5.6
0.3-0.39	880	9115	9.7
0.4-0.49	1727	9168	18.8
0.5-0.59	5015	13319	37.7
0.6-0.69	11739	18514	63.4
0.7-0.79	27686	33491	82.7
0.8-0.89	76895	82903	92.8
0.9-0.99	182490	188002	97.1
1	202514	205097	98.7

Table 4: Percentage of Russian friends to probability of being Russian

Percent Range	Russian	Total	%
0-0.09	138	19401	0.7
0.1-0.19	77	2704	2.8
0.2-0.29	93	2441	3.8
0.3-0.39	176	2380	7.4
0.4-0.49	391	2543	15.4
0.5-0.59	1476	3807	38.8
0.6-0.69	4658	6756	68.9
0.7-0.79	11371	13064	87.0
0.8-0.89	17617	18992	92.8
0.9-0.99	10512	10982	95.7
1	2978	3084	96.6

Table 5: Percentage of Canadian friends to probability of being Canadian

Percent Range	Canadian	Total	%
0-0.09	2047	179589	1.1
0.1-0.19	3547	38930	9.1
0.2-0.29	3231	8527	37.9
0.3-0.39	2677	3759	71.2
0.4-0.49	2217	2568	86.3
0.5-0.59	2408	2581	93.3
0.6-0.69	2399	2508	95.7
0.7-0.79	3053	3175	96.2
0.8-0.89	4943	5092	97.1
0.9-0.99	6657	6815	97.7
100	4321	4401	98.2

Table 6: Percentage of British friends to probability of being British

Percent Range	British	Total	%
0-0.09	1322	101244	1.3
0.1-0.19	1908	23554	8.1
0.2-0.29	1963	6722	29.2
0.3-0.39	1860	3263	57.0
0.4-0.49	1783	2488	71.7
0.5-0.59	2271	2913	78.0
0.6-0.69	2529	3088	81.9
0.7-0.79	3161	3690	85.7
0.8-0.89	5069	5539	91.5
0.9-0.99	5914	6155	96.1
1	3659	3733	98.0

Table 7: Percentage of Australian friends to probability of being Australian

Percent Range	Australian	Total	%
0-0.09	1070	108756	1.0
0.1-0.19	1350	10490	12.9
0.2-0.29	1213	2388	50.8
0.3-0.39	1040	1299	80.1
0.4-0.49	867	974	89.0
0.5-0.59	1010	1078	93.7
0.6-0.69	996	1029	96.8
0.7-0.79	1238	1289	96.0
0.8-0.89	2098	2149	97.6
0.9-0.99	3038	3115	97.5
1	1714	1741	98.4

Table 8: Percentage of Philippino friends to probability of being Philippino

Percent Range	Philippino	Total	%
0-0.09	113	15546	0.7
0.1-0.19	102	607	16.8
0.2-0.29	94	177	53.1
0.3-0.39	105	152	69.1
0.4-0.49	124	168	73.8
0.5-0.59	159	224	71.0
0.6-0.69	283	380	74.5
0.7-0.79	558	694	80.4
0.8-0.89	1051	1159	90.7
0.9-0.99	911	968	94.1
1	377	398	94.7

Table 9: Percentage of German friends to probability of being German

Percent Range	German	Total	%
0-0.09	818	61480	1.3
0.1-0.19	554	2366	23.4
0.2-0.29	337	580	58.1
0.3-0.39	252	336	75
0.4-0.49	178	247	72.1
0.5-0.59	188	250	75.2
0.6-0.69	232	302	76.8
0.7-0.79	342	404	84.7
0.8-0.89	528	574	92.0
0.9-0.99	375	393	95.4
1	128	135	94.8

Table 10: Percentage of Ukrainian friends to probability of being Ukrainian

Percent Range	Ukrainian	Total	%
0-0.09	246	31587	0.8
0.1-0.19	510	3762	13.6
0.2-0.29	575	978	58.8
0.3-0.39	565	696	81.2
0.4-0.49	486	557	87.3
0.5-0.59	517	566	91.3
0.6-0.69	529	577	91.7
0.7-0.79	535	561	95.4
0.8-0.89	542	577	93.9
0.9-0.99	362	372	97.3
1	105	107	98.1

Friend Tally Range	Correct	Incorrect	Number of Datapoints	Percent Accuracy
1 to 5	565055	80182	645237	87.57
6 to 10	272360	61647	334007	81.54
11 to 15	147453	51472	198925	74.12
16 to 20	90344	43514	133858	67.49
21 to 25	57752	35277	93029	62.08
26 to 30	38545	29478	68023	56.66
31 to 35	26408	24175	50583	52.21
36 to 40	18531	19744	38275	48.42
41 to 45	13088	16242	29330	44.62
46 to 50	9373	13371	22744	41.21
51 to 55	6942	11166	18108	38.34
56 to 60	5405	9432	14837	36.43

Table 11: Results of Applying Naive Nation Guessing System to Groupings of Users With Similar Number of Friends