Generating Look-alike Names via Distributed Representations

***Abstract*—Motivated by applications in login challenge and privacy protection, we consider the problem of large-scale construction of realistic-looking names to serve as aliases for real individuals. We seek these look-alike names to preserve name characteristics like gender, ethnicity, and frequency of occurrence while being unlinkable back to the source individual.**

We introduce the technique of *distributed name embeddings*, representing names in a high-dimensional space such that distance between name components reflects the degree of cultural similarity between these strings. We present different approaches to constructing name embeddings, and evaluate their cultural coherence. We demonstrate that name embeddings strongly encode gender and ethnicity, as well as name frequency.

I. Introduction

Names are important. The names that people carry with them are arguably the strongest single facet of their identity. Names convey cues to people's gender, ethnicity, and family history. Hyphenated last names suggest possible marital relationships. Names even encode information about age, as social trends alter the popularity of given names.

That names serve as people's primary societal identifier gives them even more power. Privacy requirements often make it undesirable or even illegal to publish people's names without their express permission. Yet there are often technical contexts where we need names which can be shared to represent things: to serve as placeholders in databases, demonstrations, and scientific studies.

In this paper, we consider the problem of constructing realistic-looking names on a large-scale to serve as aliases for real individuals. The task here is more subtle than may appear at first. We might consider randomly assigning names from reference sources such as telephone books, but these are actual names and hence violate privacy concerns. We might consider generating names at random from component first/last name parts. But these fake names will not respect gender, ethnicity, and temporal biases: consider the implausibility of names like "Wei Hernandez" or "Roberto Chen". When replacing names to anonymize a medical study, dissonance is created when female names are replaced by male ones, and the names of elderly patients aliased by newly coined names.

Our interest in generating look-alike names arose through a computer security application: how might an email user who lost their password be able to convince the account provider of their identity as part of account recovery process, or as part of the second-login challenge? We reasoned that the genuine account holder should be able to distinguish the actual contacts they have corresponded with from a background of imitation names (Fig. 1). But this is only effective when the background

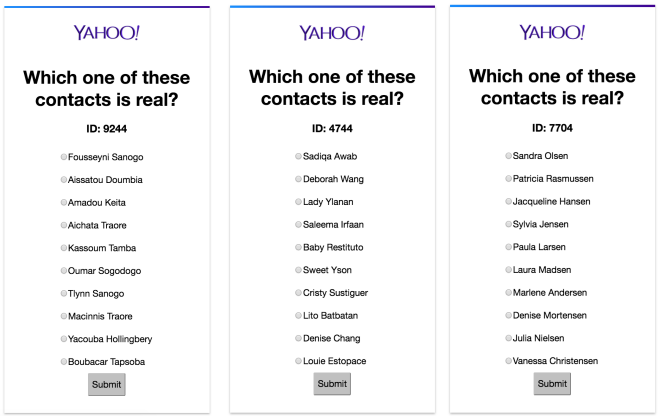

Fig. 1: Three contact list based challenges using the technique proposed in this paper. In each challenge only one contact is real.

names are culturally indistinguishable from the contacts, a property which did not hold under naive name generation methods. Otherwise, the correct answer might stand out and hence be easily guessed by an attacker who tries to take control of user accounts. For example, consider an example challenge question asked to a hypothetical user, *wendy_wong@*, given in Table I (left). If background names are generated naively without preserving ethnic properties (right), guessing the correct answer becomes much easier. However, when the generated names preserve ethnic and cultural properties of the real contacts that they replace (middle), the guessing task for attacker remains hard, because the imitation names look very similar to the real contacts.

TABLE I: A security challenge question: "pick someone you contacted among the following". Left: the contact list of a hypothetical user *wendy_wong@*. Middle: a replacement list generated using the technique proposed in this paper (retaining one real contact *Charles Wan*). Right: a naively generated random replacement list.

Real Contacts	Proposed Challenge	Naive Challenge
Angela Chiang	Amanda Hsu	John Sander
Paresh Singh	Nirav Sharma	Steve Pignootti
Charles Wan	*Charles Wan*	*Charles Wan*
Yuda Lin	Joko Yu	Jeff Guibeaux
Lin Wong	Hua Li	Sam Khilkevich
Tony Kuang	David Feng	Mary Lopez
Hua Yim	Jie Fung	Ron Clemens

The major contributions of our work are:

- *Generating realistic replacement names through name embeddings* – We propose a new technique of representing the semantics of first/last names through high-dimensional *distributed name embeddings*. By training on millions of email contact lists, our embeddings establish cultural locality among first names, last names, and the linkages between them, as illustrated by examples in Figure II. Through nearest neighbor analysis in embeddings space, we can construct replacement aliases for any given name which preserve this cultural locality.
- *Gender, racial, and frequency preservation through name embeddings* – Through computational experiments involving ground truth data from the U.S. Census and Social Security Administration, we show that our name embeddings preserve such properties as gender and racial demographics for popular names and industrial sector for corporate contacts. Even more surprisingly, our embeddings preserve frequency of occurrence, a property that to the best of our knowledge has not been previously recognized in the distributed word embedding community.
- *Establishment of ethnic/gender homophily in email correspondence patterns* – Through large-scale analysis of contact lists, we establish that there is greater than expected concentration of names of the same gender and race for all major groupings under study.

II. Building Name Embeddings

A. Methodology

We seek to construct a list of background names that looks very plausibly real, even though they are machine generated. One way to create such a list is to start from real contact names, and replace each one with a look-alike names of the same gender, ethnicity and name frequency. However this approach requires multiple complex components, including reliable ethnicity and gender classifiers. Instead we propose to do aways with these complex steps by deriving the signals directly from a large amount of data, through name embedding.

In our approach, each name, first or last, is embedded in high dimensional space as a high dimensional vector using word2vec [1]. Our hypothesis (verified in Section III-C) is that people have a tendency to contact people of the same ethnicity and gender. Consequently, when using the contact lists of millions of users as a text corpus, the resulting embedding of names would capture this tendency by placing names of the same gender and ethnicity close-by in the high-dimensional space. For each real name in the contact list, we can then choose a name near it in the high dimensional space. The resulting replacements should have good gender and ethnicity similarity to the original names (Table I (middle)). Furthermore, it turns out that in aggregate, the name frequency of the look-alike names also resemble that of the real names very well.

B. Data Sources and Preparation

In this section, we introduce the datasets that are used in this study, as well as a detailed description about our data preparation process.

Data Sources. Datasets employed in our work are:

- *Contact Lists*– This set of data, here after referred to as the *contact lists* is a proprietary sample of contact lists from 2 million distinct Yahoo email users. The length of each contact list varies from 1 to 21, because longer lists have been truncated. Names in each contact list are ordered by contacting frequency/recency. A much larger set of contacts for around 1 billion Yahoo users are available to us and the results using this larger set is very similar to those from the smaller set. Therefore in this paper we present all our results based on the smaller set. Each entry of a list is a full name but not necessarily a human name. To preserve the privacy of users, the owner associated with each contact list was not available in the data.
- *Census 1990* [2]– The Census 1990 dataset is a public dataset from US Census website. It records the frequently occurring surnames from US Census 1990. This dataset contains 4,725 popular female names and 1,219 popular male names.
- *Census 2000* [3] – The Census 2000 dataset is another public dataset from US Census website. It contains the frequently occurring 151,672 surnames from US Census 2000. Associated with each name is a distribution over six categories of races. The races are: White, Black, Asian/Pacific Islander (API), American Indian/Alaskan Native (AIAN), Two or more races (2PRACE), and Hispanics. In this paper we refer to the races and ethnicity interchangeably.

The ethnicity distributions of *Census 2000* and *Contact lists* data are given in Table III.

Data Preparation. The *contact lists* data records the social interaction of email users. These contact lists include substantial noise in the name fields. Artifacts include omitted names [4], sometimes marked by an arbitrary string like "zzzzzzzz", or names that are meaningful but clearly not human, like "Microsoft", "Facebook", and "GEICO". Moreover, many names in contact lists data are partial, with only the first name or the last name present.

C. Word2vec Embeddings

The word2vec software [5] is an efficient tool to learn the distributed representation of words for large text corpus. It comes with two models: the Continuous Bag-of-Words model (CBOW) and the Skip-Gram (SG) model. The CBOW model predicts the current word based on the context while the Skip-Gram model does the inverse and maximizes classification of a word based on another word in the same context [6].

We start our analysis by using the cleaned *contact lists* and the word2vec software [5]. Each contact list is treated as a sentence, and together they form a text corpus. Unless

Male names	1th NN	2nd NN	3rd NN	4th NN	5th NN	Female names	1th NN	2nd NN	3rd NN	4th NN	5th NN
Andy	Andrew	Ben	Chris	Brian	Steve	Adrienne	Allison	Aimee	Amber	Debra	Amy
Dario	Pablo	Santiago	Federico	Hernan	Diego	Aisha	Aliyah	Nadiyah	Khadijah	Akil	Aliya
Elijah	Isaiah	Joshua	Jeremiah	Bryant	Brandon	Brianna	Brittany	Briana	Samantha	Jessica	Christina
Felipe	Rodrigo	Rafael	Eduardo	Fernando	Ricardo	Candy	Connie	Becky	Angie	Cindy	Christy
Heath	Brent	Chad	Brad	Brett	Clint	Chan	Wong	Poon	Ho	Wai	Yip
Hilton	Xooma	Eccie	Erau	Plexus	Gapbuster	Cheyenne	Destiny	Madison	Brittany	Taylor	Kayla
Isaac	Samuel	Israel	Eli	Esther	Benjamin	Dominque	Renarda	Lakenya	Lakia	Lashawna	Shatara
Jamal	Jameel	Kareem	Anmar	Khalifa	Nadiyah	Ebonie	Lakeshia	Tomeka	Ebony	Latasha	Shelonda
Lamar	Terrell	Derrick	Eboni	Tyree	Willie	Florida	Fairfield	Integrity	Beacon	Southside	Missouri
Mohammad	Shahed	Mohmmad	Ahmad	Rifaat	Farishta	Gabriella	Daniella	Vanessa	Marilisa	Isabella	Elisa
Moshe	Yisroel	Avraham	Gitty	Rivky	Zahava	Giovanna	Giovanni	Elisa	Paola	Giuliana	Mariangela
Rocco	Vito	Salvatore	Vincenza	Pasquale	Nunzio	Han	Jin	Yong	Sung	Huan	Teng
Salvatore	Pasquale	Nunzio	Gennaro	Vito	Tommaso	Kazuko	Keisuke	Junko	Yumi	Yuka	Tomoko
Thanh	Minh	Thuy	Thao	Ngoc	Khanh	Keren	Ranit	Galit	Haim	Zeev	Rochel

TABLE II: The five nearest neighbors(NN) of representative male and female names in embedding space, showing how they preserve associations among Asian (Chinese, Korean, Japanese, Vietnamese), British, European (Spanish, Italian), Middle Eastern (Arabic, Hebrew), North American (African-American, Native American, Contemporary), and Corporate/Entity.

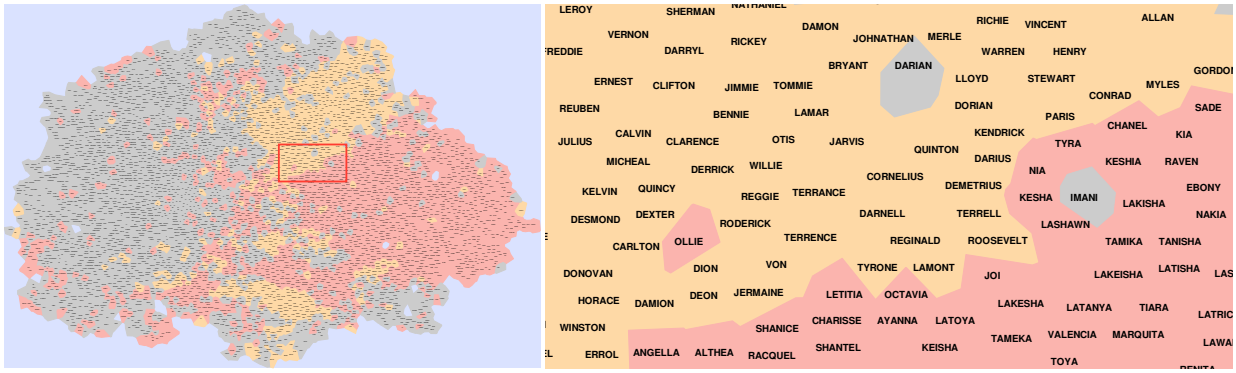

Fig. 2: Visualization of the name embedding for the most frequent 5,000 first names from email contact data, showings a 2D projection view of name embedding (left). The pink color represents male names while orange denotes female names. Gray names have unknown gender. The right figure presents a close view along the male-female border, centered around African-American names.

	Census 2000		*Contact lists*	
Races	Count	Percentage	Count	Percentage
White	115,167	0.8593	12,837,406	0.7428
Black	5,262	0.0393	544,983	0.0315
API	6,100	0.0455	1,323,888	0.0766
AIAN	268	0.002	19,272	0.0011
2PRace	131	0.001	7,934	0.0005
Hispanics	7,089	0.0529	2,548,329	0.1475
Total	134,017		17,281,812	

TABLE III: Ethnicity distribution of *Census 2000* data and that of intersected names between the *contact list* and *Census 2000* data. The label information comes from *Census 2000*.

otherwise stated, all results in the paper are based on the CBOW model with the default word2vec parameter settings (see Section II-D for comparison of different models and ways of constructing the corpus). The output of word2vec is a dense matrix of dimension $517,539 \times 100$, with each unique name represented as a row of the matrix.

Embedding Visualization. To understand the landscape of the name embeddings, we visualize the names as a 2D map. We used the stochastic neighborhood embedding [7] to reduce the original 100-dimensional embedding to 2D. We assign each name to a cluster using gender/ethnicity ground truth, and created the maps using `gvmap` [8].

Figure 2 (left) illustrates the landscape of first names. This visualization establishes that the embedding places names of the same gender close-by. Using Census data, we color male names orange, female names pink, and names with unknown gender gray. Overall names of the same gender form mostly contiguous regions, indicating that the embedding correctly capture gender information by placing names of the same gender close-by. Figure 2 (right) is an inset showing a region along the male/female border. We can see that "Ollie", which is considered a predominantly female name per Census data (2:1 ratio of female/male instances), is placed in the male region, close to the male/female border. Per [9], we found that "Ollie" is more often a male name, and used as a nickname for "Oliver" or "Olivia". Hence our embedding is correct in placing it near the border. The embedding also correctly placed "Imani" and "Darian", two names not labelled by the Census data, near the border, but in the female/male regions, respectively. Per [9], "Imani" is a African name of Arabic origin, and can be both female and male, mainly female; "Darian" can also be female and male, but mainly male, and

is a variant of "Daren" and "Darien", among others.

Fig. 3 (left) presents a map of the top 5000 last-names. We color a name according to the dominant racial classification from the Census data. The top 5000 names contain four races: White (pink), African-American (orange), Hispanic (yellow), and Asian (green). Names without a dominant race are colored gray. The three cutouts in Fig. 3 highlight the homogeneity of regions by cultural group. The embedding clearly places White, Hispanic and Asian in large contiguous regions. African-American names are more dispersed. Interestingly, there are two distinct Asian regions in the map. Fig. 4 presents insets for these two regions, revealing that one cluster consists of Chinese names and the other Indian names. Overall, Fig. 2 and Fig. 3 show that our name embeddings capture gender and ethnicity information well.

D. Evaluation of Different Word2vec Embeddings

The embedding from word2vec is influenced by two factors: the input text, and the word2vec parameter settings. So far we have been using the *contact lists* unchanged as the input to word2vec. However there are many possible variants. Since certain contact lists are short, and word2vec uses a sliding window, some names from the contact lists of the previous/next persons could be in the same window as that for the current person. One possible solution is to pad stop words in between two contact lists such that a sliding window would never enclose contacts from unrelated persons. Instead of putting both first and last names together in one embedding space, another variant might generate two embeddings, one using only the first names, and another using the last names. In addition to the input text, the second factor that influences the embedding is the different settings of models and parameters, for example, the selection between CBOW model and SG model, the size of sampling window and the size of negative samples.

To understand how these two factors influence the embedding, in particular with regard to the quality of the resulting look-alike names, we evaluate the following variants of the word2vec embeddings:

- Set the word2vec model to be CBOW or SG.
- Generating joint embeddings of first names and last names using the contact lists as they are ("CBOW joint" or "SG joint").
- Generating embedding for first names and last names separately by including only first/last names in the contact lists ("CBOW sep" or "SG sep").

Metrics. To evaluate the quality of the embeddings with regard to look-alike names, we propose three metrics to measure popularity, gender and ethnicity similarities between real and look-alike names.

Firstly, we would like the overall frequency of a name in the real contact lists to be similar to that in the look-alike lists. For example, if "Mark" is a more popular name than "Barnabas" in the real contact lists, we would also expect that "Mark" appears more often as a replacement name than "Barnabas". We define two types of frequencies. The *real name frequency* is the frequency of names in the *Contact list*. The *replacement usage frequency* is the frequency of a name in the replacement name population. To measure the popularity dis-similarity, we sample 10k names randomly from the name list, with sampling probability proportional to *real name frequency*. We record ten nearest neighbors (NN) for each of them. Now the popularity dis-similarity is computed by *Jensen-Shannon Divergence* using the *real name frequency* of the sampled names and *replacement usage frequency* of the nearest neighbor names. Secondly, we measure gender similarity by precision at k, defined as the percentage of the k-nearest neighbors of a name having the same gender as the name itself. Finally, we measure ethnicity similarity by precision at 1. For example, the precision for White names is defined as $P(W|W) = P(\text{1st NN is White}|\text{real name is White})$.

Results. We present the evaluation results in Table IV. The table shows that a joint embedding of first and last names are often better than the separate embedding. In addition, the CBOW model generally outperformed the SG model for the majority of the nearest neighbor tests. Given these observations, in this paper by default we use "CBOW joint".Note that while $P(B|B)$ (35%-59%) is generally much lower than $P(W|W)$ (92%-94%), considering that a randomly picked name from the contact list has a probability of 74% of being White but only a probability of 3% of being Black, $P(B|B)$ is actually significantly above the probability of a random name being black.

III. Properties of Name Embeddings

After all the names in the *contact lists* have been embedded through word2vec, each name is represented by a vector in the high dimensional space. Earlier, in Fig. 2 and Fig. 3, we have provided visual evidence that the embedding is coherent, in the sense that it places names of similar gender and ethnicity close-by. In Section II-D we have also seen aggregate numerical evidence of this coherence. In this section we evaluate the coherence of the name embedding quantitatively and in detail.

A. Gender Coherence and Analysis

We first examine the gender coherence of a subset of first names and their ten nearest neighbors. This subset of first names is the intersection between *contact lists* and *Census 1990*. It contains 1,146 unique male first names and 4,009 unique female first names. All names in the subset are ranked by their popularity as measured in the *Census 1990* data. Table V shows the gender coherence results, measured by precision at k, as a function of the population of the names, and k, the number of nearest neighbors. For example, the cell at $\{\leq 20\%, 2\}$ of Table V (left) reads 97. It means that for the top 20% most popular names, 97% of their nearest 2-neighbors have the same gender as them. To save space, we only report the first two significant digits of each precision (e.g., 0.9742 becomes 97). In addition we color the cells of the tables based on the values. Within each table, we use warm colors for high values and cold color for low values. This gives us heat-maps through which it is easier to see the trend of how the precision

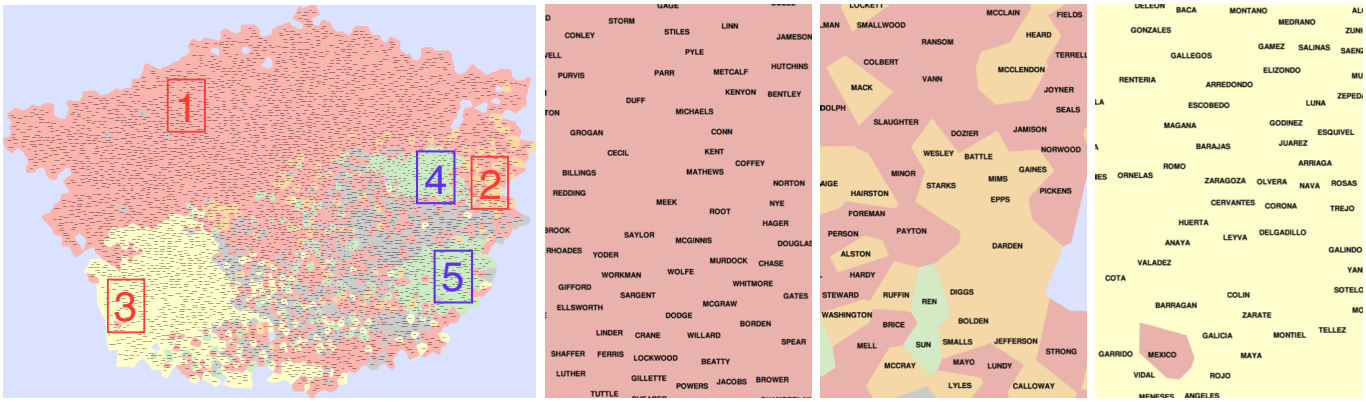

Fig. 3: Visualization of the name embedding for the top 5000 last names, showings a 2D projection view of the embedding (left). Insets (left to right) highlight British 1, African-American 2 and Hispanic 3 names.

Variation	Popularity	Gender		Ethnicity (NN(1))			
		NN(1)	NN(10)	$P(W\|W)$	$P(B\|B)$	$P(A\|A)$	$P(H\|H)$
CBOW joint	0.6434(0.0007)	0.9092	**0.9360**	0.9362	**0.5939**	**0.7626**	**0.7543**
SG joint	0.6747(0.0002)	0.8844	0.9274	**0.9461**	0.4561	0.7208	**0.7543**
CBOW sep	0.6675(0.0003)	**0.9162**	0.9350	0.9299	0.4437	0.7167	0.6710
SG sep	**0.5776(0.0001)**	0.8844	0.9205	0.9217	0.3451	0.6797	0.6971

TABLE IV: Evaluation of different embedding variants. *Popularity* is measured by Jensen-Shannon Divergence of frequency distribution, while all other values are precision at k.

varies with popularity of the first name, and the number of neighbors.

From Table V, we observe that our proposed name embedding scheme shows strong gender coherence, especially for popular names. As we can see from the tables, the percentage of neighbors that have same gender as the original first name is very high for the top 30% most popular names comparing to a randomly assigned name (50%). On the other hand, the percentage decreases when unpopular names are included, and also decreasing as the number of neighbors increases.

A similar coherence analysis was carried out with regard to ethnicity. We found that the top neighbors of a popular name tend to have a high probability of being in the same ethnicity group. The coherence for an ethnic group correlates positively with the popularity of the group in the *contact lists*. The coherence for AIAN and 2PRACE are poor, because they only account for 0.1% and 0.05% of the last names in the contact lists. Thus there is too little data to get the embedding correct.

Top %	1	2	3	4	5	6	7	8	9	10
≤ 10%	100	99	98	98	98	98	98	98	98	98
≤ 20%	99	97	96	96	95	95	95	95	95	95
≤ 30%	96	95	94	93	93	92	93	92	92	91
≤ 40%	93	93	91	90	90	89	89	89	88	88
≤ 50%	89	89	86	85	85	84	84	84	83	83
≤ 60%	86	86	84	83	82	82	82	81	81	80
≤ 70%	82	81	79	79	78	77	77	76	76	76
≤ 80%	79	78	76	75	74	74	74	73	73	72
≤ 90%	76	75	73	73	72	71	71	71	70	70
All	73	72	70	69	69	68	68	67	67	67

Top %	1	2	3	4	5	6	7	8	9	10
≤ 10%	97	97	97	96	95	95	95	95	95	95
≤ 20%	91	91	91	90	89	89	89	89	88	88
≤ 30%	85	84	84	84	83	83	82	82	82	81
≤ 40%	80	79	79	78	78	77	77	77	76	76
≤ 50%	75	74	74	73	73	72	72	72	71	71
≤ 60%	69	69	68	68	67	67	66	66	66	66
≤ 70%	66	65	66	65	64	64	63	63	63	63
≤ 80%	62	61	61	61	60	60	59	59	59	59
≤ 90%	59	59	59	58	58	57	57	57	57	57
All	57	56	56	56	55	55	55	54	54	54

TABLE V: Gender coherence of the name embedding for males (left) and females (right), as measure by the percentage of k-neighbors being male (left) and female (right).

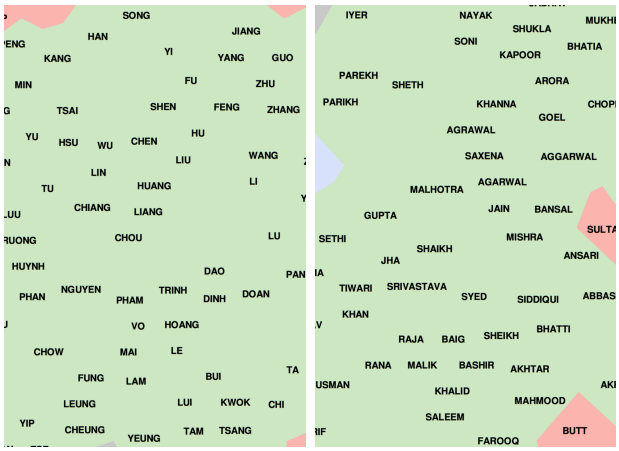

Fig. 4: The two distinct Asian clusters. Left: Chinese/South Asian names (4 in Fig. 3). Right: Indian names (5 Fig. 3).

	PCC		SCC	
	RA	RU	RA	RU
First names	0.5813	0.7795	0.5170	0.5402
Last names	0.2260	0.4454	0.3444	0.3916

TABLE VI: Correlation of real names and replacement names.

B. Name Popularity Analysis

To measure the popularity preserving of word embedding, we calculate the *real name frequency* of a name (R), the average *real name frequency* of its replacement names (ten nearest

neighbors) (A), and its *replacement usage frequency* (U). Two measurements, Pearson's correlation coefficient (PCC) and Spearman's rank correlation coefficient (SCC), are used to measure how well the popularity is preserved. The results are shown in Table VI. Overall we can see that the correlation between the *real name frequency* R and the *replacement usage frequency* U is higher than that for the *real name frequency* R and its neighbors' *real name frequency* A. This indicates that a popular name is very likely to appear in among the nearest neighbors of other names, even though its nearest neighbors are not necessarily popular names.

C. Why does the embedding work?

Recall that we start from contact lists of millions of users, and embedding the first and last names using word2vec. We have observed visually and though numerical metrics that the embedding captures gender and ethnicity very well. A natural question is: why does it work so well?

Our hypothesis is that the population as a whole have a bias toward contacting people of the same ethnicity and gender. This bias drives the embeddings of names of similar ethnicity and gender close to each other, and pushes names of different ethnicity and gender apart. To test this hypothesis, we look at the frequency distribution of percentage of males in our mailing list, and compare with the null hypothesis. In Fig. 5 (left), we divide contact lists by a threshold based on the minimum number T of identifiable genders in the list. E.g., $T = 5$ means those contact lists with at least five gender-identifiable names. The distributions of the ratio of identifiable males in the contact lists with $T = 5$ and 10 are seen as the two lower curves. Clearly, the majority of the contact list has around 50% males. However, looking at these distribution along would not tell us whether the distributions have any bias. For this purpose, we need to compare them with the null hypothesis.

We generate the null distribution by assigning the gender of a name randomly following the gender distribution of the *contact list*. Fig. 5 (left) shows that the distributions based on the null hypothesis (the two higher curves) are spikier, with around 30% of the contact lists having 50% of males, compared with the observed 15%. Fig. 5 (right) shows the deviation of the observed distribution from the null hypothesis. It shows a clear bimodal pattern, confirming our hypothesis that contact lists on average exhibit a bias towards either male domination, or female domination, especially the latter. A similar analysis was done with regard to ethnicity, and confirmed a bias for people to contact other people of the same ethnicity. Due to space limition details are ommitted.

To further verify the gender bias in observed contact lists, we model the observed number of males in all contact lists as a Binomial mixture model. The mixture model fits the observed data well, and suggests that 47% of the observed contact lists belong to male users and 53% belongs to female users.

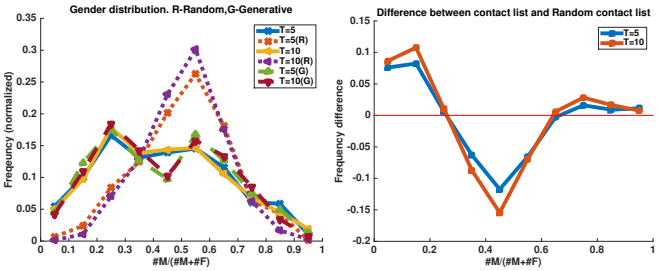

Fig. 5: Left: the distribution of user's gender in contact lists data. Distributions with legend "R" are from binomial distribution with probability 0.5. Distributions with legend "G" are from binomial mixture model with parameters inferred using EM algorithm. Other distributions are from observation in the *contact lists*. Right: deviation from the null hypothesis.

IV. Conclusion

We propose a new technique for generating look-alike names through distributed name embeddings. By training on millions of email contact lists, our embeddings establish gender and cultural locality among names. The embeddings make possible construction of replacement aliases for any given name that preserve gender and cultural identity. Through large-scale analysis of contact lists, we establish that there is a greater than expected concentration of names of the same gender and race for all major groupings under study. The look-alike name generation technique is currently being deployed as part of soft-challenge pipeline for login challenge and account recovery see Fig. 1). A demostration website for generation look-alike names is available as http://yo/lookalikename.

We are current using these embeddings for age and ethnicity classification of names. Gender results are available in the URS datahub and ethnicity results will be in soon.

References

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[2] C. Bureau, https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html, 1990.

[3] ——, https://www.census.gov/topics/population/genealogy/data/2000_surnames.html, 2000.

[4] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*. ACM, 2005, pp. 71–80.

[5] open source project, https://code.google.com/archive/p/word2vec/, 2013.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of Workshop at ICLR*, 2013.

[7] L. Van Der Maaten, "Accelerating t-sne using tree-based algorithms," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245, 2014.

[8] Y. Hu, E. Gansner, and S. Kobourov, "Visualizing graphs and clusters as maps," *IEEE Computer Graphics and Applications*, vol. 30, pp. 54–66, 2010.

[9] M. Campbell, http://www.behindthename.com, 1996.