T. Swearingen and A. Ross. "A label propagation approach for predicting missing biographic labels in face-based biometric records," IET Biometrics, Vol. 7, Issue 1, pp. 71 - 80, 2018.

A Label Propagation Approach for Predicting Missing Biographic Labels in Face-Based Biometric Records

Thomas Swearingen and Arun Ross

***Abstract*—A biometric system uses the physical or behavioral attributes of a person, such as face, fingerprint, iris or voice, to recognize an individual. Many operational biometric systems store the biographic information of an individual, *viz.,* name, gender, age and ethnicity, besides the biometric data itself. Thus, the biometric record pertaining to an individual consists of both biometric data and biographic data. In this work, we propose the use of a graph structure to model the relationship between the biometric records in a database. We show the benefits of such a graph in deducing biographic labels of incomplete records, *i.e.,* records that may have missing biographic information. In particular, we use a label propagation scheme to deduce missing values for both binary-valued biographic attributes (*e.g.,* gender) as well as multi-valued biographic attributes (*e.g.,* age group). Experimental results using face-based biometric records consisting of name, age, gender and ethnicity convey the pros and cons of the proposed method.**

I. Introduction

Biometrics is the process of recognizing individuals based on their physical or behavioral attributes by using automated or semi-automated methods [1]. Examples of such attributes include face, fingerprint, iris, voice, gait and signature. A typical biometric system acquires the biometric data of an individual (*e.g.,* a face image) and stores it in a database along with an identifier (*e.g.,* the name of the individual). The data corresponding to an individual constitutes the *biometric record* of that individual. Thus, the database or *gallery* of a biometric system contains multiple biometric records pertaining to multiple individuals.

In some biometric applications, the biometric record of an individual may be supplemented with additional biographic data (such as name, age, gender, ethnicity and occupation) or social network data (such as friends in FaceBook or connections in LinkedIn). For example, the UIDAI Aadhaar program in India,[1] the OBIM program in the United States,[2] the TWIC program in the United States,[3] and the E-VERIFY program in the United States[4] collect the biographic details of an individual besides their biometric data. In such applications, the biometric record of an individual in the gallery will contain both biometric and biographic data.

Typically, the gallery records are viewed as *independent* entities. For example, in a biometric identification system, the input probe data (*e.g.,* an unknown face image) is *independently* compared against each gallery record in order to determine the identity of the probe. While in some cases the gallery data may be automatically clustered into multiple categories (*e.g.,* see [3]), in general, the relationship between the gallery records is seldom modeled or exploited in the biometric recognition process.

In this work, we consider the use of a simple graph to model the relationship *between* gallery records. Each node in the graph would correspond to a biometric record and each edge weight would denote the similarity between two biometric record (nodes). Thus, the gallery would be denoted as a graph where a connection between two nodes indicates the degree of similarity between two identities. Figure 1 illustrates such a graph. The use of such a graph to model the relationship between gallery records has several advantages:

1) The output of the identification process would be a subgraph consisting of not only "matching" candidates whose face images look similar to the probe image, but also other candidate images that are "related" to the probe. For example, when searching for the identity of an unknown probe image in the graph, the output

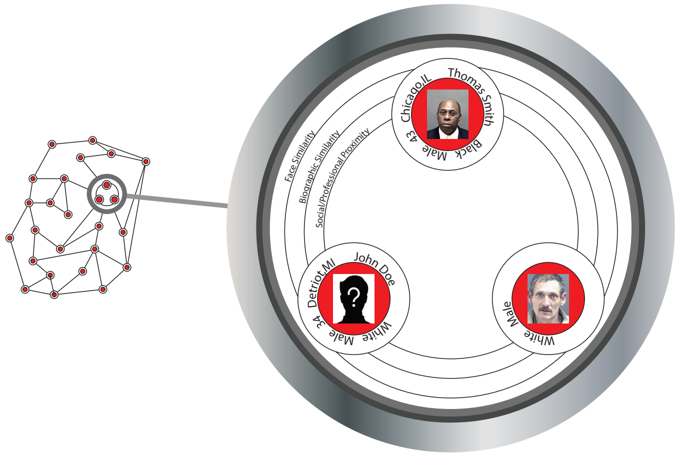

Fig. 1. An example graph where each node in red represents a person and each edge represents the similarity between two people. This similarity is a function of the biometric information, biographic information and, potentially, social media information (but not in this work). Face images are from the MORPH database [2].

T. Swearingen and A. Ross are with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, 48824 USA e-mail: swearin3@msu.edu, rossarun@cse.msu.edu.

[1] https://uidai.gov.in

[2] http://www.dhs.gov/obim

[3] https://www.tsa.gov/for-industry/twic

[4] http://www.uscis.gov/e-verify

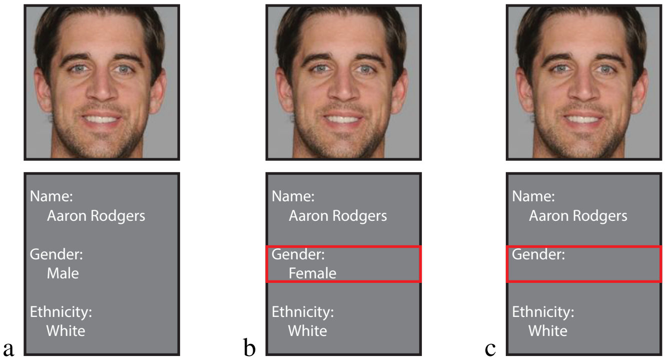

Fig. 2. Examples of various types of errors that can occur in biometric records of identity management systems: (a) Complete and Correct Record; (b) Complete but Erroneous Record; (c) Incomplete Record.

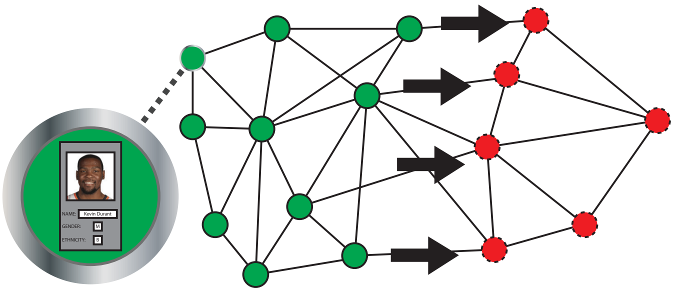

Fig. 3. A graph where each node represents a record and each edge weight represents the similarity between a pair of nodes. Green nodes with a solid border indicate complete records and red nodes with a dotted border represent incomplete records. The goal of our work is to induce a set of biographic labels to the incomplete nodes based on labels in the complete nodes.

TABLE I
RELATED WORKS WHICH PREDICT GENDER FROM FACE IMAGES. EACH WORK APPROACHES THE PROBLEM FROM A DIFFERENT PERSPECTIVE: GALLAGHER AND CHEN [5] COMBINE BIOMETRIC AND BIOGRAPHIC INFORMATION, SHAN [6] USES A TRADITIONAL BIOMETRIC PIPELINE (*i.e.,* FACE IMAGE → FEATURE EXTRACTION → CLASSIFICATION), AND LEVI AND HASSNER [7] USE A DEEP LEARNING METHOD.

Work	**Approach**	**Dataset**	**Data Size**	**Accuracy (%)**
Gallagher and Chen [5]	Biometrics+ Biographics	Proprietary	148 images	81.7
Shan [6]	Traditional Biometric Pipeline	LFW	13,233 images	94.81
Levi and Hassner [7]	Deep Learning	Adience	19,487 images	86.8 ± 1.4

may consist of gallery identities that are in social or professional proximity to the individual (such as a close friend or a co-worker). This would help in cases where the identity of the probe is not in the gallery, but related identities are present in the graph.

2) When a node, pertaining to the biometric record of an individual, is incomplete (*e.g.,* missing demographic data), the graph structure can be leveraged to predict the missing information if necessary. Further, the graph structure can be utilized to detect nodes that may have incorrect demographic information.

In this work, we will explore one aspect of this graph structure. We will demonstrate the possibility of using such a graph structure to deduce missing biographic data in a record. Some nodes are likely to have missing or incorrect biographic labels (*e.g.,* see Figure 2). This is especially true since many of the identity management systems mentioned earlier have high collection rates. For example, the UIDAI Aadhaar project collects 15 million records per month and US TWIC collects 30,000 records per month. The incoming data is likely to contain a variety of typographical errors, selection errors (Figure 2b), or missing values (Figure 2c). The rapid rate of data collection may preempt the possibility of manually reviewing each biometric record for accuracy. An automated method may, therefore, be required to verify data in a biometric record (*i.e.,* node).

Consider a graph in which we have a set of nodes that have correct and complete information (complete nodes) and another set of nodes where a specific attribute (*e.g.,* gender) has incorrect, missing, or unverified information (incomplete nodes). In our graph-based system, highly similar nodes (*i.e.,* those with high edge weight values between them) are more likely to have similar biographic profiles. We can leverage this similarity to induce a labeling on the incomplete nodes by "pushing" labels from the complete nodes. Figure 3 illustrates this concept. As shown in the figure, each node in the graph is a record that includes the face image, name, gender, and ethnicity of an individual. There is a set of complete nodes (represented in green) and a set of incomplete nodes (represented in red). The goal here is to induce labels on the red nodes using the similarity information between **all the nodes** (both green and red). We utilize a label propagation method to accomplish this.

This work is an extension of our previous work [4]. In the current work, we extend our previous approach as follows: (a) the label propagation method is used to predict multi-valued biographic attributes rather than just binary-valued attributes; and (b) we explore the utility of different attributes in the prediction process by assigning different weights to individual attributes when computing the similarity between nodes during label propagation. For comparison, we use existing automated methods that can predict gender, ethnicity, and age from a single face image.[5]

Section II provides a review of related literature. Section III details the baseline methods for predicting biographic information from faces and names. Section IV presents our proposed method for predicting biographic information using the graph-based gallery and label propagation. Section V reports the experiments and results. In Section VI, we analyze the results. Section VII provides a summary of the paper.

[5] The goal of this paper is **not** to develop a better gender or ethnicity or age-group classifier. Rather, the goal is to demonstrate that label propagation using a graph-based representation of the gallery is a viable way to impute information to incomplete nodes. Consequently, such an approach can be used in the future to "predict" biographic or demographic labels for which classifiers are not available (*e.g.,* occupation).

II. Related Work

A. Predicting Biographic Information

The term "soft biometrics" is often used to refer to attributes of an individual that cannot, in isolation, be used for distinctively recognizing a person but, which, in conjunction with primary biometric attributes, such as face or fingerprint, can help improve the identification accuracy of a biometric system [8], especially in challenging environments [9]–[11]. Examples of such traits include gender, age, ethnicity, *etc*. Thus, the biographic data of an individual could be viewed as soft biometric attributes. Often, such soft biometric information can be automatically gleaned from the primary biometric data (*e.g.,* age and gender from face images). On a smaller scale, there have been attempts to predict a person's occupation or name from a face image [12], [13], but with relatively less success. Dantcheva *et al.* [14] provide a comprehensive survey on the topic of soft biometrics.

Gallagher and Chen [5] propose estimating gender from face images using a probabilistic model of faces and first names. Shan proposed a method for gender prediction from face images using Adaboost to select the best Local Binary Pattern (LBP) features which were then classified as Male or Female using a SVM classifier [6]. Levi and Hassner [7] use a Convolutional Neural Network (CNN) to learn features and classify the gender of face images. Makinen and Raisamo [21] provide a survey of gender estimation techniques. Table I summarizes these efforts.

In the problem of age estimation, there are two categories: (a) age classification and (b) age regression. In age classification, a face image is assigned to one pre-defined age group. In age regression, the precise age is estimated given a face image. Gallagher and Chen [5] estimate age from face images using a probabilistic model of faces and first names. Levi and Hassner [7] propose a CNN approach. Han *et al.* [15] use Biologically-Inspired Features (BIF) and a hierarchical classifier to estimate precise age from a face image. The hierarchical classifier consists of a series of SVM classifiers which split the face images into groups pertaining to particular age ranges. For each group, a regressor is learned which outputs a final precise age value. Chen *et al.* [16] extend the hierarchical approach to use CNNs to separate the images into groups as well as perform the final age regression. Fu *et al.* [22] provide a comprehensive survey on age estimation from face images. Table II summarizes some of these works.

The problem of ethnicity prediction is a difficult one. Given that the most-common problem is to predict a person's ethnicity from a face image, the vast majority of automated ethnicity predictors base their prediction on facial appearance. Ding *et al.* [17] extract local texture features and shape features from 3D and 2D face images to predict ethnicity labels. Kumar *et al.* [19] pair color histograms with an SVM classifier to predict ethnicity labels. Muhammad *et al.* [18] explore the use of Weber Local Descriptors (WLD) and Local Binary Patterns (LBP) for ethnicity prediction. Ambekar *et al.* [20] develop an ethnicity predictor based on names using hidden Markov models and decision trees. Fu *et al.* [23] provide a survey on ethnicity estimation from face images. Table III summarizes these works.

B. Combining Biographic Information and Biometrics

Biographic information, such as gender or ethnicity, may be used by an automated recognition system to help facilitate matching [30]. There are two different strategies to integrate biographic data into a biometric system: (a) the biographic information can be used to filter the gallery database such that the input probe is only compared against those gallery records sharing a similar biographic profile [24], [25] and (b) the biometrics and biographics are combined at the match score level in order to improve the recognition accuracy [26]–[28]. The importance of fusing biometric and biographic data has been acknowledged by commercial enterprises as well [31], [32].

Klare *et al.* [24] showed that using biographic-specific matchers can improve the identity retrieval performance. The authors tested a face recognition system on a variety of different cohorts (specific values of a biographic attribute, *e.g.,* Male or Female for the gender attribute) and found that face recognition systems performed better on some cohorts compared to other cohorts. Specifically, the matchers had difficulty recognizing the Female, Black, and Younger (18 to 30 years old) cohorts. They also showed that the recognition performance on a specific cohort increased if the matcher was trained only on images from the same cohort.

Han *et al.* [25] describe a sketch-to-photo face matching scheme that uses gender information to filter a gallery of mugshot images. They found that the matching performance increased if probe sketch images were matched to only those mugshot images in the gallery having the same gender as the probe.

Jain *et al.* [8] proposed a scheme to combine soft biometric information (gender, ethnicity, height) with the fingerprints of an individual using a Bayesian scheme. The proposed method was observed to improve the recognition performance of the fingerprint matcher.

Tyagi *et al.* [27] use a likelihood ratio-based fusion method to combine the match scores emerging from the biometric matcher and the biographic matchers. They test their method on a synthetic dataset consisting of fingerprint images from the NIST-BSSR1 dataset and names and addresses from a database of electoral records. They found that this resulted in better recognition accuracy then when the biometric classifier and the biographic classifiers were used separately.

Bhatt *et al.* [28] combine biometric and biographic match scores for a de-duplication application. A match score is computed between each corresponding attribute in two records. The attributes considered in their work are fingerprint, name, father's name, and address. This comparison between two records results in a 4-dimensional match vector. A SVM is then trained to differentiate between training samples labeled as 'duplicate' (-1) and 'non-duplicate' ($+1$). The data is synthetically generated from four fingerprint datasets (CASIA fingerprint V5, MCYT, WVU multi-modal, FVC 2006) and two unnamed biographic datasets. The output of the SVM

TABLE II
Related works which predict age from face images only.

Work	Source Attribute	Dataset	Dataset Size	Age Ranges	Performance Metric	Performance
Gallagher and Chen [5]	Face	Proprietary	148 images	Age Regression	Mean Absolute Error	9.33
Levi and Hassner [7]	Face	Adience	19,487 images[†]	0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60+	Classification Accuracy	50.7% $\pm$ 5.1%
Han *et al.* [15]	Face	FG-NET	1,002 images	Age Regression	Mean Absolute Error	3.8 ± 4.2
		MORPH II	78,207 images			3.6 ± 3.0
		PCSO	100,012 images			4.1 ± 3.3
Chen *et al.* [16]	Face	Adience	26,580 images[†]	0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60+	Classification Accuracy	52.88% $\pm$ 6%
		FG-NET	1,002 images	Age Regression	Mean Absolute Error	3.49
		Chalearn Challenge	4,699 images	Age Regression	Gaussian Error	0.297

[†] While both works use the same dataset, it appears that Levi and Hassner did not use the entire Adience dataset.

TABLE III
Related works which predict ethnicity.

Work	Source Attribute	Dataset	Dataset Size	Ethnicity Classes	Performance Metric	Performance
Ding *et al.* [17]	Face	FRGC v2.0	4,007 faces	Asian, Non-Asian	Classification Accuracy	98.26%
		BU-3DFE	2,500 faces	White, Asian	Classification Accuracy	97.88%
Muhammad *et al.* [18]	Face	FERET	2,368 faces	Asian, Black, Hispanic, Asian-Middle-Eastern, White	Classification Accuracy	96%
Kumar *et al.* [19]	Face	PubFig	58,797 faces	Asian, Black, Indian, White	Classification Accuracy	94.6%
Ambekar *et al.* [20]	Name	Wikipedia	127,596 Names	Greater European, Greater African, Asian, Greater East Asian, Western European, African, British, East Asian, Eastern European, French, German, Hispanic, Indian Sub-Continent, Italian, Japanese, Jewish, Muslim, Nordic	F-Score	0.69 (Average)

TABLE IV
Works which combine biographic data and biometrics.

Work	Dataset(s)	Biometric Attribute(s)	Biographic Attribute(s)	Use of Biographic Info
Klare *et al.* [24]	Pinellas County Sheriff's Office (PCSO)	Face	Gender, Race, Age	Filter Gallery
Han *et al.* [25]	AR Face Database Pinellas County Sheriff's Office (PCSO)[†] Multiple Encounter Dataset II (MEDS-II)[†]	Face, Face Sketch	Gender	Filter Gallery
Jain *et al.* [26]	Proprietary	Fingerprint	Gender, Ethnicity, Height	Fused with Biometric Score
Tyagi *et al.* [27]	NIST–BSSR1	Fingerprint	Name, Address	Fused with Biometric Score
Bhatt *et al.* [28]	CASIA fingerprint V5 MCYT WVU multi-modal FVC 2006	Fingerprint	Name, Father's Name, Address	Fused with Biometric Score
Sudhish *et al.* [29]	Proprietary	Face, Fingerprint	Name, Father's Name	Fusion with Biometric Score

[†] This dataset was used as a background dataset.

is a score indicating the distance of the match vector from the margin. This score is then used to make a decision as to whether two records are duplicates or not.

Sudhish *et al.* [29] also combine biometric and biographic matchers for de-duplication. They use an adaptive fusion scheme with multiple biographic attributes *and* multiple biometric attributes. The fusion scheme adapts to use different attributes based on information available, computational cost, and desired accuracy. They test their method on a synthetic dataset created from face images from the PCSO dataset, fingerprint images from NIST Special Database 14, and biographic information from the US census.

In previous literature, the role of the name attribute in the context of biometrics has not been adequately addressed. Social context influences the decision of choosing an appropriate name for an individual based on factors such as gender and ethnicity [13]. Therefore, the name can reveal information about an individual's other biographic data. For instance, Liu and Ruths [33] used first names as features to predict gender in Twitter. Ambekar *et al.* [20] proposed the use of hidden Markov models and decision trees to classify names into different cultural/ethnic groups. Moreover, some other works have explored the connection between names and faces. Chen *et al.* [13] demonstrated, on a small scale, that first names of face images can be predicted at rates greater than chance.

C. Label Propagation

Label propagation, a type of semi-supervised learning method, uses both labeled and unlabeled data. This differs from supervised learners that utilize labeled data only or unsupervised learners that work with unlabeled data. There are two types of semi-supervised classifiers: (1) transductive learners and (2) inductive learners. Inductive learners allow for data to be added after completion of the training stage, while transductive learners require all data to be available at the training stage.

The label propagation method proposed by Zhou *et al.* [34], falls into the transductive learner category. Their method first constructs a fully-connected graph of all of the data points (nodes). The similarity between pairs of nodes is found using a Gaussian Radial Basis Function. The labels are then propagated from the labeled nodes to the unlabeled nodes according to a loss function with a normalized Laplacian which promotes labeling with local and global consistency (*i.e.,* both nodes that are close in the feature space (local) and nodes which lie in the same structure or manifold (global) are likely to have the same label).

The fundamental assumption of Label Propagation is that points that are likely to have the same label lie on the same manifold. The goal is to induce labels on the unlabeled data using the labeled points and the underlying manifold in the data. Label Propagation has been used in a variety of application such as image segmentation [35], image annotation [36], and recommender systems [37]. One particularly relevant problem where label propagation has been applied is to improve the labeling in datasets where there are missing or incorrect labels [38], [39].

III. Traditional Approach: Predicting Biographic Information from a Single Attribute/Record

In this work, we consider predicting three biographic attributes: gender, ethnicity, and age-group. Gender is simple to understand as it is typically assumed to take one of two values: Male or Female.[6] Age is another biographic attribute that is simple to understand as it is just the number years since an individual's birth. In this work, we discretize the age value into 3 age groups: 29 and under, 30-44, and 45 and over. Ethnicity is much more complex as it is often a group-defined construct that can change over time. While there are commonly defined cohorts (*e.g.,* Asian, Black, *etc.*), these cohorts fail to accurately reflect the heterogeneity of each group [40]. This has led to difficulty in tracking these groups for many social scientists [41]. In our work, we use three ethnicity labels: White, Black, and Hispanic. However, the proposed method can be easily expanded to other labeling schemes.

The thrust of this work is in harnessing a label propagation method which uses (a) multiple attributes in a record and (b) the relationship between these attributes, to predict a missing biographic attribute. For comparison to existing work, we will also predict the biographic attribute using a single attribute (*e.g.,* face) in a record. These single attribute predictors are described in the following subsections. In particular, we predict gender and ethnicity from the name only; and gender, ethnicity, and age from the face image only.

A. Deducing Gender and Ethnicity from Names

Before we describe the label propagation method used in this work, we first establish *baseline* methods where gender or ethnicity are deduced from a single record. In this regard, below we describe algorithms for deducing gender/ethnicity from names or face images.

1) Names to Gender Database (NGD): C't, a German computing magazine, published a database of 47,780 names and their corresponding gender labels [42]. It includes 20,288 Male names, 19,181 Female names, and 8,311 Unisex names. The names are from 54 countries which are classified as *Male/Female/Unisex* by native speakers of the language. We refer to this database as the Names-to-Gender Database (NGD). Here, we treat the unisex label as "unknown." A lookup table is used to classify an input name as Male, Female, or unknown. Figure 4 shows an overview of this method.

2) USCB-1990 Database: The United States Census Bureau (USCB) undertook a project to determine undercount following the 1990 Decennial Census. This project amassed 6.3 million usable census records that included names of people. In 1995, the USCB published a summary of this information for genealogical reasons [43]. The summary includes three files, each of which contains four fields: name, frequency in percent, cumulative frequency in percent, and rank. The three files correspond to *Male* forenames, *Female* forenames, and *all* surnames. Note that forenames and surnames are not linked.

[6] However, in many contexts gender can take on more than two values (*e.g.,* http://www.cnn.com/2014/02/13/tech/social-media/facebook-gender-custom/index.html).

We utilize the Male and Female forenames files to create a forename-based gender classifier.

The classifier is developed as follows. The likelihood for a given forename, $\Pr(N = n \mid G)$, can be computed based on the frequencies provided in these files, as long as the forename occurs in them. Here, N is the name variable, n is a specific name, and G is the gender variable (which can take on values M or F). If a forename does not occur in one of the files, then $\Pr(N = n \mid G) = 0.0$ for that particular name. The posterior probability of a forename being Male or Female can then be calculated as follows. First, we note that $\Pr(N = n) = \Pr(N = n \mid G = M) + \Pr(N = n \mid G = F)$. In the USCB-1990 dataset, there are 6,188,353 different *people* whose gender is given, of which, 3,184,399 are Female and 3,003,954 are Male.[7] Thus, the prior probabilities are set to $\Pr(G = F) = 0.515$ and $\Pr(G = M) = 0.485$. The posterior probability is then computed as:

$$\Pr(G \mid N = n) = \frac{\Pr(N = n \mid G)\Pr(G)}{\Pr(N = n)}. \tag{1}$$

For a given forename, both $\Pr(G = F \mid N = n)$ and $\Pr(G = M \mid N = n)$ are calculated, and the forename is assigned to that category whose posterior probability is the largest. To avoid division by zero, if $\Pr(N = n \mid G = M) + \Pr(N = n \mid G = F) = 0$, then we set the posteriors of both the Male and Female class to 0.5. If the posterior probabilities are equal, then the gender of the forename is classified as "unknown". Figure 5 shows an overview of this method.

3) USCB-2000 Database: In order to provide the general public with genealogical, marketing, and cultural research tools, the United States Census Bureau (USCB) published a list of surnames and their corresponding ethnicity distributions [44]. The report uses the responses from the approximately 270 million people counted during the 2000 Decennial Census. The USCB distilled the responses into a set of 151,671 surnames. The ethnicity-wise percentage for each surname was made available, with the caveat that some percentages were obscured to assure confidentiality.[8] Only surnames with more than 100 occurrences were reported to assure confidentiality.

The ethnicity categories available in the USCB-2000 Database are Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian/Pacific Islander, Non-Hispanic American Indian/Alaskan Native, Non-Hispanic of 2 or more Races, and Hispanic Origin. In this work, we summarize this information into four classes (White, Black, Hispanic, and Unknown). Thus, given a surname we compute the posterior probabilities for each of the four classes, *i.e.*, $\Pr(E = B \mid N = n)$, $\Pr(E = H \mid N = n)$, $\Pr(E = W \mid N = n)$ and $\Pr(E = U \mid N = n)$, where E represents the ethnicity variable which can take on values Black (B), Hispanic (H), White (W), and Unknown (U).

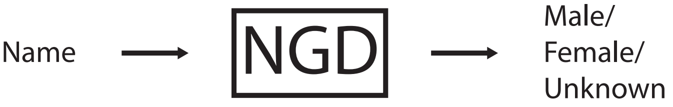

Fig. 4. The Names-to-Gender (NGD) Database is used to map an input name to a gender label.

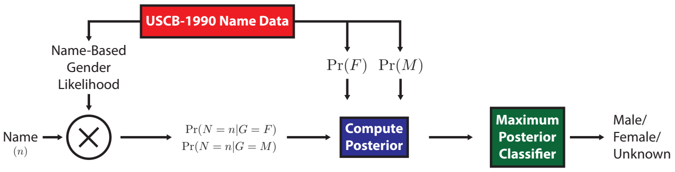

Fig. 5. Overview of the USCB-1990 Gender-from-Name Classifier. A forename, n, is input into the system and a gender label, {*Male, Female, Unknown*}, is output.

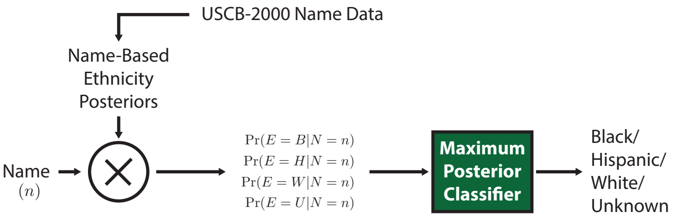

Fig. 6. Overview of the USCB-2000 Ethnicity-from-Name Classifier. A surname, n, is input into the system and an ethnicity label, {*Black, Hispanic, White, Unknown*}, is output.

The unknown class is an agglomeration of the Non-Hispanic Asian/Pacific Islander, Non-Hispanic American Indian/Alaskan Native, Non-Hispanic of 2 or more Races classes in the database. If a surname is not present in the database, then the probability of all classes is set to 0.0. The surname is classified based on the maximum posterior probability rule. If the posterior probabilities of a surname are equal, then the surname is classified as unknown. Figure 6 shows an overview of this method.

B. Biographic Prediction from Face Image

In this work, we used a Commerical-Off-the-Shelf (COTS) system to predict age, ethnicity, and gender from face images. The COTS system takes a face image as input and outputs an ethnicity probability for each of the following categories: White, Black, Asian, Hispanic, or Other. We label a face image with the ethnicity corresponding to the largest probability. Since in this work we only consider three ethnicity classes, White, Black and Hispanic, the Asian and Other labels are interpreted as "unknown." The software also outputs a Male or Female probability which we use to determine a gender label based on the larger probability. Lastly, the software outputs an age value in years. We use this value to assign an age group label to the face in one the following three age ranges: 29 & under, 30–44, or 45 & older.

[7] There are 4,275 *unique* Female forenames and 1,219 *unique* Male forenames.

[8] In the case where percentages are suppressed for some ethnicities corresponding to a particular surname, we sum the percentages that are available, subtract it from 100%, and divide it evenly among the suppressed percentages for that particular surname.

IV. Proposed Approach: Predicting Biographic Information Using Multiple Identity Records

A. Label Propagation

Unlike the single attribute based approaches mentioned above, we now predict biographic attributes using all of the available attributes. In addition, the proposed method uses evidence from multiple records to predict biographic attributes. We first construct a graph where each node corresponds to a biometric record and each edge weight value defines the similarity between nodes (records). In this graph, there are two types of nodes:

1) **Complete Node:** A nodal record which has no missing/incorrect fields.
2) **Incomplete Node:** A nodal record that has one or more missing/incorrect biographic fields.

We use a label propagation method to push labels from the complete nodes to the incomplete nodes [34]. Suppose that we have n records, v of which are complete and $n-v$ of which are incomplete. We represent this as $\mathcal{R} = \{R_1, \ldots, R_v, R_{v+1}, \ldots, R_n\}$. We first construct a label matrix $\mathbf{Y} \in R^{n \times d}$ where d is the number of biographic cohorts. For example, when predicting ethnicity which has labels Black, Hispanic and White, then $d = 3$. Each row in $\mathbf{Y}$ corresponds to a node in the graph. The first v rows of $\mathbf{Y}$ correspond to the complete nodes and the last $n-v$ nodes correspond to the incomplete nodes. The first v rows of $\mathbf{Y}$ are all zeros except for a single 1 in the column corresponding to the label of that node. The last $n-v$ rows of $\mathbf{Y}$ are all zeros as there is no label for the incomplete nodes. For example, suppose we have four nodes, two complete and two incomplete. The first complete node has the ethnicity label Black and the second has the ethnicity label White. The label matrix $\mathbf{Y}$, where column 1 corresponds to "Black", column 2 corresponds to "Hispanic", and column 3 corresponds to "White", looks like:

$$\mathbf{Y} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

For this formulation, each biographic attribute is comprised of discrete, finite-valued labels. The set of labels is given by $\mathcal{L} = \{0, 1, \ldots, d-1\}$. In general, let the set $\{y_1, y_2, \ldots, y_v\}$, where $y_i \in \mathcal{L}$, denote the biographic labels of the complete nodes.

Algorithm 1 describes the biographic label propagation method. The record set, $\mathcal{R}$, the label matrix, $\mathbf{Y}$, the attribute weights, $\mathcal{B}$, and two parameters, σ and α, are taken as input. We first must calculate the affinity matrix for the graph which is done by comparing each record. The $f_{\text{diff}}(R_i, R_j)$ function on Line 6 returns a scalar value indicating the difference between records R_i and R_j. Further details of record comparison are given in Section IV-B. The affinity matrix is then normalized with the sum of each row which yields the similarity matrix $\mathbf{S}$. The label matrix $\mathbf{Y}$ is then used to let label information "flow" from complete nodes to incomplete nodes. This "flow" is facilitated by the node relationships manifested as values in $\mathbf{S}$.

As the original authors noted, we can compute the final values directly rather than iteratively pushing label information [34]. This is accomplished using the $\mathbf{F}^* = (\mathbf{I} - \alpha\mathbf{S})^{-1}\mathbf{Y}$ function. The $(\mathbf{I} - \alpha\mathbf{S})^{-1}$ part of the function can be viewed as a diffusion kernel which diffuses the complete node labeling from the upper (complete) section of $\mathbf{Y}$ to the lower (incomplete) section of $\mathbf{Y}$. For a particular node (*i.e.,* a specific row in $\mathbf{Y}$ and $\mathbf{F}^*$), label information is collected in each column of $\mathbf{F}^*$. The larger the value in a particular column, the more likely an incomplete node belongs to the class corresponding to that column. Continuing our previous ethnicity example, $\mathbf{F}^*$ will have three columns. Suppose row i corresponds to an incomplete node, if $F^*_{i,0} > F^*_{i,1}$ and $F^*_{i,0} > F^*_{i,2}$ then incomplete node i is predicted to have label value 0, replacing the existing value in a record or populating the missing field.

Algorithm 1 Biographic Label Propagation

```
 1: procedure PropagateLabels($\mathcal{R}, \mathbf{Y}, \mathcal{B}, \sigma, \alpha$)
 2:     for $i, j \in [1, n]$ do
 3:         if i = j then
 4:             $W_{ij} = 0$
 5:         else
 6:             $W_{ij} = \exp\left(-\frac{f_{\text{diff}}(R_i, R_j, \mathcal{B})^2}{2\sigma^2}\right)$
                                        ▷ Edge weights are based on record similarity.
 7:         end if
 8:     end for
 9:     $D_{ii} = \text{zeros}(n)$
10:     for $i \in [1, n]$ do
11:         $D_{ii} = \sum_{j=1}^{n} W_{ij}$
                    ▷ Diagonal entries are the sum of the corresponding row in $\mathbf{W}$.
12:     end for
13:     $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$
14:     $\mathbf{F}^* = (\mathbf{I} - \alpha\mathbf{S})^{-1}\mathbf{Y}$          ▷ $\mathbf{F}^*$ is the same size as $\mathbf{Y}$
15:     for $i \in (v, n]$ do
16:         $l_i = \text{argmax}_{0 \leq j < k} F^*_{ij}$
17:     end for
18:     return $l_i$'s                              ▷ Labels for incomplete nodes.
19: end procedure
```

B. Record Comparison Techniques

Name: We use levenshtein distance to compare names. The distance is normalized to $[0, 1]$ range by dividing the levenshtein distance by the length of the longest string. Thus, $\phi_{\text{n}}(R_i, R_j)$ returns a value between 0 and 1 indicating the distance between the name fields in record R_i and record R_j.

Face: We use a commercial-off-the-shelf (COTS) face matcher to compare face images. The COTS matcher returns a similarity score in the $[0, 1]$ range. This score is transformed to a distance score by subtracting the similarity score from 1. Thus, $\phi_{\text{f}}(R_i, R_j)$ returns a value between 0 and 1 indicating the distance between the face images of record R_i and record R_j.

Age, Ethnicity, Gender: These attributes have a finite number of values (*e.g.,* Male or Female for gender). The distance is 0 if the values in the two records are the same and 1 if the values are different. Thus, $\phi_{\text{a}}(R_i, R_j)$, $\phi_{\text{e}}(R_i, R_j)$, and $\phi_{\text{g}}(R_i, R_j)$ return a value between 0 and 1 indicating the distance between the age, ethnicity, and gender fields, respectively, in record R_i and record R_j.

Combining the Attributes: The $f_{\text{diff}}(R_i, R_j, \mathcal{B})$ function compares two records. Here, $\mathcal{B}$ denotes the set of weights.

A distance score is computed for each *available* attribute. All attributes may not be available for each record and will be ignored if not available. For example, when predicting the gender, the gender attribute may not be available for the incomplete records. The $f_{\text{diff}}(R_i, R_j, \mathcal{B})$ function summarizes these distance scores into a single value by taking a weighted average. If R_i and R_j are both complete records, then $f_{\text{diff}}(R_i, R_j, \mathcal{B})$ is given by

$$f_{\text{diff}}(R_i, R_j, \mathcal{B}) = \frac{1}{|\mathcal{A}|} \sum_{k \in \mathcal{A}} \beta_k \, \phi_k(R_i, R_j), \tag{2}$$

where $\mathcal{A} = \{n, f, a, e, g\}$ denotes the various attributes (name, face, age, ethnicity, or gender) in the record, β_k denotes the weight of attribute k, and $|\mathcal{A}|$ is the cardinality of the set $\mathcal{A}$ (5 in this case). The set of weights $\mathcal{B} = \{\beta_n, \beta_f, \beta_a, \beta_e, \beta_g\}$ is used, as it is possible that some attributes are more important than others and, therefore, the attributes should be weighted differently. Each beta value can take on a value between 0 and 1. If either R_i or R_j are incomplete records, and continuing with the example of predicting gender, then $f_{\text{diff}}(R_i, R_j, \mathcal{B})$ is given by:

$$f_{\text{diff}}(R_i, R_j, \mathcal{B}) = \frac{1}{|\mathcal{A} \setminus g|} \sum_{k \in \mathcal{A} \setminus g} \beta_k \, \phi_k(R_i, R_j). \tag{3}$$

The gender attribute (represented by g) is removed from the set of attributes $\mathcal{A}$ under consideration for any comparison that includes at least one incomplete record. Similarly, the ethnicity (e) or age (a) attribute would be ignored if either R_i or R_j are incomplete records, and we were predicting the ethnicity or age attribute.

V. Experiments

A. Dataset

A typical face dataset includes face image(s) of multiple subjects and occasionally includes biographic information such as gender, ethnicity, or age. Rarely do such datasets include names of subjects. Some datasets that are comprised of celebrities contain names (such as LFW [45]). However, none of the datasets include *all* of these attributes, *viz.*, faces image, name, gender, ethnicity, and age. Therefore we assembled our own dataset based on images from the Web called the Knox County Arrest Dataset (KCAD). Unlike the work by Tyagi *et al.* [27], Bhatt *et al.* [28], Sudhish *et al.* [29] that use synthetic datasets, our work utilizes real naturally occurring datasets. This dataset is an expanded version from our earlier work [4].

The Knox County Sheriff's Office (KCSO) posts the information of arrestees every 24 hours. This information contains the arrestee's: name, gender, ethnicity, age, and face mugshot. We compiled this information for use in our experiments. The number of records is given in Table V as well as a breakdown by biographic attribute. In order to avoid the class imbalance problem, when predicting an attribute, we will only use the number of records from each class equal to the number of records from the smallest class. For example, when predicting gender, we use 2,322 Female records and 2,322 Male records even though there are 5,712 Male records available.

TABLE V
Biographic Details of the Knox County Arrest Dataset (KCAD).

Attribute	Cohort	Number of Records
Gender	Male	4984
	Female	2019
Ethnicity	Black	1522
	Hispanic	154
	White	5299
	Other	28
Age	29 & Younger	2476
	30-44	3166
	45 & Older	1361
Total		7003

TABLE VI
Results of biographic prediction via label propagation using all attributes, equally weighted.

Attribute	σ	α	Cohort	Mean Acc. ± STD
Age Group	0.11	0.03	≤ 29	$76.5\% \pm 0.916\%$
			30-44	$50.5\% \pm 1.05\%$
			≥ 45	$76.0\% \pm 1.94\%$
			Overall	$67.7\% \pm 0.744\%$
Ethnicity	0.14	0.02	Black	$88.2\% \pm 6.04\%$
			Hispanic	$74.4\% \pm 3.14\%$
			White	$59.4\% \pm 4.35\%$
			Overall	$73.9\% \pm 2.47\%$
Gender	0.1	0.01	Male	$95.6\% \pm 0.577\%$
			Female	$91.8\% \pm 1.34\%$
			Overall	$93.7\% \pm 0.925\%$

B. Biographic Prediction

1) Label Propagation Using All Attributes, Equally Weighted: Section IV-A details the Label Propagation method used to predict gender, ethnicity, and age group. For each attribute prediction, we use 4-fold cross-validation. In this experiment, we use equal weights for all attributes such that the weights sum to 1 (*i.e.*, $\beta_k = 0.2 \, \forall \, k \in \{n, f, a, e, g\}$). We also perform a parameter search to find the best value of σ and α. This is a two-stage process: first, we vary both σ and α from 0.1 to 0.9 in increments of 0.1. Once we find the best values for σ and α, we do another parameter search, in 0.01 increments starting at the best value from the first stage minus 0.09 going to the best value from the first stage plus 0.09 (*e.g.,* if the best value from the first parameter search is 0.6, then the second parameter search would vary from 0.51 to 0.69 in increments of 0.01).

The results of age group prediction are given in Table VI. There are 1,361 records with age 29 & under, 1,361 records with age 30-44, and 1,361 records with age 45 & older in four folds. The fields used are: name, face image, age group, gender, and ethnicity. All fields are used when comparing complete records, and the name, face, ethnicity and gender fields are used when comparing two incomplete records or a complete record and an incomplete record.

The results of ethnicity prediction are given in Table VI.

TABLE VII
RESULTS OF BIOGRAPHIC PREDICTION VIA LABEL PROPAGATION ON A SUBSET OF ATTRIBUTES, EQUALLY WEIGHTED. IN THE "FACE ONLY" RUN, ONLY THE FACE ATTRIBUTE IS USED TO COMPARE RECORDS (*i.e.,* $\beta_f = 1.0$ AND $\beta_k = 0.0 \,\forall\, k \in \{n, a, e, g\}$). IN THE "BIOGRAPHIC ONLY" RUN, ONLY THE BIOGRAPHIC ATTRIBUTES ARE USED TO COMPARE RECORDS (*i.e.,* $\beta_f = 0.0$ AND $\beta_k = 0.25 \,\forall\, k \in \{n, a, e, g\}$).

	Attributes Used	σ	α	**Cohort**	**Mean Acc. ± STD**
AGE GROUP	Face Only	0.2	0.04	≤ 29	87.6% ± 1.87%
				30-44	43.9% ± 0.565%
				≥ 45	89.6% ± 2.00%
				Overall	73.7% ± 0.990%
	Biographic Only	0.07	0.09	≤ 29	34.2% ± 1.14%
				30-44	32.0% ± 1.88%
				≥ 45	58.7% ± 1.31%
				Overall	41.6% ± 0.895%
ETHNICITY	Face Only	0.27	0.33	Black	96.1% ± 3.95%
				Hispanic	75.6% ± 5.88%
				White	89.7% ± 2.65%
				Overall	87.1% ± 2.66%
	Biographic Only	0.1	0.05	Black	53.6% ± 1.08%
				Hispanic	74.4% ± 1.81%
				White	41.3% ± 3.77%
				Overall	56.5% ± 1.43%
GENDER	Face Only	0.2	0.01	Male	96.5% ± 0.495%
				Female	98.3% ± 1.15%
				Overall	97.4% ± 0.752%
	Biographic Only	0.3	0.35	Male	42.5% ± 1.22%
				Female	80.2% ± 3.19%
				Overall	61.4% ± 1.55%

TABLE VIII
ATTRIBUTE WEIGHTS LEARNED THROUGH A TWO-STAGE PARAMETER SEARCH.

Attribute	β_n	β_f	β_a	β_e	β_g	σ	α
Age Group	0.0	0.4	0.6	0.0	0.0	0.09	0.16
Ethnicity	0.7	0.3	0.0	0.0	0.0	0.41	0.3
Gender	0.5	0.1	0.0	0.0	0.4	0.17	0.07

There are 154 White records, 154 Black records, and 154 Hispanic records in four folds. The fields used are: name, face image, age group, gender, and ethnicity. All fields are used when comparing complete records, and the name, face, age group and gender fields are used when comparing two incomplete records or a complete record and an incomplete record.

The results of gender prediction are given in Table VI. There are 2,019 Male records and 2,019 Female record in four folds. The fields used are: name, face image, age group, gender, and ethnicity. All fields are used when comparing complete records, and the name, face, ethnicity and age group fields are used when comparing two incomplete records or a complete record and an incomplete record.

2) Label Propagation Using A Subset of Attributes, Equally Weighted: In order to measure the importance of the biometric attribute compared to the biographic attributes, the label propagation method is first executed on a graph whose edge weights are based *only* on the face score and then executed on another graph whose edge weights are computed *without* the face score (*i.e.,* biographic attributes only). That is, for one run $\beta_f = 1.0$ and $\beta_k = 0.0 \,\forall\, k \in \{n, a, e, g\}$ and for the other run $\beta_f = 0.0$ and $\beta_k = 0.25 \,\forall\, k \in \{n, a, e, g\}$. The results of age group, ethnicity, and gender prediction are given in Table VII. The values for σ and α are determined using the same search scheme described in Section V-B1.

3) Label Propagation Using Learned Weights: It is possible that some attributes are more important than others. To find the best value for the weights, we vary the set of weights ($\mathcal{B} = \{\beta_n, \beta_f, \beta_a, \beta_e, \beta_g\}$) as well as σ and α. We use a two-stage parameter search to find the best set of weights for the data for each prediction problem (age group, ethnicity, gender). We first vary the weights from 0.0 to 1.0 in 0.1 step increments with the constraint that the sum of the weights must be 1. In addition, we vary the σ and α parameters in increments of 0.1. Second, we use the best values from the first stage and then vary *only* the σ and α parameters in 0.01 step increments starting at the best value from the first stage minus 0.09 going to the best value from the first stage plus 0.09 (*e.g.,* if the best value from the first parameter search is 0.6, then the second parameter search would vary from 0.51 to 0.69 in increments of 0.01). Each attribute produced a different set of values for $\mathcal{B}$, as is shown in Table VIII.

The results of age group prediction are given in Table IX with $\beta_n = 0$, $\beta_f = 0.4$, $\beta_a = 0.6$, $\beta_e = 0$, and $\beta_g = 0$. The run time was 14 minutes, 25 seconds (0.64 seconds/record). The fields used are: name, face, age group, gender, and ethnicity. All fields are used when comparing complete records, and the name, face, ethnicity and gender fields are used when comparing two incomplete records, or a complete record and an incomplete record.

The results of ethnicity prediction are given in Table IX with $\beta_n = 0.7$, $\beta_f = 0.3$, $\beta_a = 0$, $\beta_e = 0$, and $\beta_g = 0$. The run time was 17 seconds (0.11 seconds/record). The fields used are: name, face, age group, gender, and ethnicity. All fields are used when comparing complete records, and the name, face, age group and gender fields are used when comparing two incomplete records or a complete record and an incomplete record.

The results of gender prediction are given in Table IX with $\beta_n = 0.5$, $\beta_f = 0.1$, $\beta_a = 0$, $\beta_e = 0$, and $\beta_g = 0.4$. The run time was 13 minutes, 20 seconds (0.40 seconds/record). The fields used are: name, face, age group, age, gender, and ethnicity. All fields are used when comparing complete records, and the name, age group, ethnicity and face fields are used when comparing two incomplete records or a complete record and an incomplete record.

4) Baseline Biographic Prediction: Section III details the methods that can predict a biographic attribute based on a single attribute of a person (*e.g.,* name or face). Ethnicity can be predicted based on surname using the USCB-2000 method or based on face using the COTS system. Age can be predicted based on the face using the COTS system. Gender can be predicted based on forename using the USCB-1990 method or NGD method, or based on the face using the COTS system.

TABLE IX
RESULTS OF BIOGRAPHIC PREDICTION VIA LABEL PROPAGATION USING LEARNED WEIGHTS.

	β_n	β_f	β_a	β_e	β_g	σ	α	**Cohort**	**Mean Acc. ± STD**
AGE GROUP	0.0	0.4	0.6	0.0	0.0	0.09	0.16	≤ 29	$85.3\% \pm 1.21\%$
								30-44	$52.9\% \pm 0.789\%$
								≥ 45	$87.0\% \pm 1.65\%$
								Overall	$75.1\% \pm 0.888\%$
ETHNICITY	0.7	0.3	0.0	0.0	0.0	0.41	0.3	Black	$98.0\% \pm 2.18\%$
								Hispanic	$87.2\% \pm 4.80\%$
								White	$89.7\% \pm 1.82\%$
								Overall	$91.6\% \pm 1.87\%$
GENDER	0.5	0.1	0.0	0.0	0.4	0.17	0.07	Male	$98.2\% \pm 0.567\%$
								Female	$98.2\% \pm 0.485\%$
								Overall	$98.2\% \pm 0.456\%$

TABLE X
RESULT OF BIOGRAPHIC PREDICTION USING BASELINE METHODS.

	Source Attribute	**Method**	**Cohort**	**Mean Acc. ± STD**
AGE GRP.	Face	COTS	≤ 29	$72.2\% \pm 1.28\%$
			30-44	$79.0\% \pm 1.52\%$
			≥ 45	$66.0 \pm 1.94\%$
			Overall	$72.4\% \pm 0.879\%$
ETHNICITY	Surname	USCB-2000	Black	$11.1\% \pm 2.11\%$
			Hispanic	$70.5\% \pm 5.29\%$
			White	$91.6\% \pm 3.30\%$
			Overall	$58.0\% \pm 2.06\%$
	Face	COTS	Black	$89.5\% \pm 6.19\%$
			Hispanic	$75.6\% \pm 2.87\%$
			White	$99.4\% \pm 1.11\%$
			Overall	$88.1\% \pm 2.31\%$
GENDER	Forename	NGD	Male	$80.1\% \pm 0.995\%$
			Female	$69.8\% \pm 1.39\%$
			Overall	$75.0\% \pm 0.907\%$
	Forename	USCB-1990	Male	$87.8\% \pm 1.01\%$
			Female	$86.6\% \pm 1.18\%$
			Overall	$87.2\% \pm 0.275\%$
	Face	COTS	Male	$99.8\% \pm 0.0857\%$
			Female	$91.7\% \pm 0.918\%$
			Overall	$95.7\% \pm 0.477\%$

TABLE XI
RESULTS OF HYPOTHESIS TEST BETWEEN LABEL PROPAGATION ACCURACIES AND COTS ACCURACIES.

Attribute	$\mathbf{a}_l - \mathbf{a}_c$	**p-value**	**Result**
Age Group	$[3.53, 4.80, 0.39, 2.06]$	0.125	Fail to Reject H_0
Ethnicity	$[7.76, 0.86, 1.72, 3.45]$	0.125	Fail to Reject H_0
Gender	$[3.07, 1.98, 2.17, 2.67]$	0.125	Fail to Reject H_0

The results are given in Table X.

5) Comparison of COTS and Label Propagation Methods: We compare the results of the COTS performance with the best label propagation method (label propagation with learned weights) to determine if the methods achieve similar performance. We have 4 accuracies for each method where each accuracy comes from one of the four folds. We represent the accuracies from the label propagation method as the vector $\mathbf{a}_l$ where the first entry is the accuracy of the label propagation method on first fold, the second entry is the accuracy of the label propagation method on second fold, *etc*. Similarly, we represent the accuracies from COTS as the vector $\mathbf{a}_c$. As the accuracies are paired, we use the Wilcoxon signed-rank test [46] to compare the accuracies. For this test,

$$\begin{aligned}\text{H}_0 &: \mathbf{a}_l - \mathbf{a}_c \text{ comes from a distribution with } 0 \text{ median}\\ \text{H}_1 &: \mathbf{a}_l - \mathbf{a}_c \text{ comes from a distribution with median}\\ &\qquad\qquad \text{different than } 0\end{aligned}$$

The results are shown in Table XI.

VI. ANALYSIS

In Section V-B1, we saw that age and gender prediction using the all attribute, equal weight label propagation method had comparable performance to the baseline classifiers (in Section V-B4). However, the overall ethnicity prediction accuracy was $\sim 15\%$ lower compared to the baseline. This is because the label propagation method was far worse at predicting White records compared to the baseline. The label propagation method had similar performance to the baseline when predicting Black and Hispanic records.

In Section V-B2, we observed that for all three prediction problems (age, ethnicity, gender), label propagation using only the face match scores (*i.e.,* $\beta_f = 1.0$ and $\beta_k = 0.0 \, \forall \, k \in \{n, a, e, g\}$) is better than using strictly the biographic attributes only – name, gender, ethnicity, and age match scores (*i.e.,* $\beta_f = 0.0$ and $\beta_k = 0.25 \, \forall \, k \in \{n, a, e, g\}$). Ethnicity prediction had a substantial increase in accuracy when using only the face match scores, compared to ethnicity prediction in Section V-B1 (+13.2%). This indicates that face match scores play a critical role in the propagation of biographic labels – which is intuitive as the face is the most discriminative of all the attributes. The biographic only accuracy was low for all three attributes, but it was still above random chance (33.3%

for age group and ethnicity, 50% for gender) in each case. This indicates that propagation with only the biographic attributes is not appropriate, but it is possible when combined with other attributes it could add additional predictive value.

In Section V-B3, we saw that label propagation using the learned weights had the best performance for all three attributes (Age Group, Ethnicity, and Gender). We also saw in Section V-B5 that the difference between the accuracies of the label propagation method with the learned weights and the COTS method were statistically insignificant for all attributes. **However, the goal of this work is not to achieve state-of-the-art biographic prediction performance, but to show the benefits of utilizing a graph-structure to model gallery records.**

The set of weights ($\mathcal{B}$) had different values when predicting different attributes. For age group, $\beta_f = 0.4$ and $\beta_a = 0.6$, while the other weights are 0.0. This is very intuitive as age group information is obviously useful when predicting the age group and the face attribute is the most discriminative attribute. For ethnicity and gender, understanding the weight values is less intuitive. For ethnicity, all β's are 0.0 except for $\beta_f = 0.3$ and $\beta_n = 0.7$ (Table VIII). Like age group, the face attribute is useful for prediction. However, based on the results of the weights of age group prediction, we would expect that the ethnicity weight (β_e) would be important for ethnicity prediction, but that is not true. Instead, the name attribute is important. This could be because there are many subjects in the dataset with the same surname who all have the same ethnicity (*e.g.,* in the 2000 U.S. Census, 98.1% of people with the surname "Yoder" reported being White [44]). The weights for gender prediction are similar to both age group prediction and ethnicity prediction as the weights for the face attribute, name attribute, and gender attribute are non-zero (see Table VIII). The face attribute and gender weight value are intuitive to understand as gender information is obviously useful and the face attribute is the most discriminative attribute. The NGD and USCB-1990 classifiers predict gender from forename with 75.0% and 87.2% prediction accuracy, respectively. This indicates that forename is a strong indicator of gender. If two subjects have the same forename, the output of $\phi_n(R_i, R_j)$ will be lower for these two records. Since forename is an indicator of gender, having the same forename (and thus a lower $\phi_n(R_i, R_j)$) is indicative of having the same gender. Although there are obviously some forenames which are gender-ambiguous (*e.g.,* Oakley[9]), gender-ambiguous names are likely less common than gender specific names.

Age group prediction via label propagation had good prediction accuracy for the ≤ 29 and ≥ 45 cohorts, but much lower performance for the $30 - 44$ cohort (for all label propagation weight schemes). This indicates that method can separate the records in a general sense (*i.e.,* younger/older), but is not good at delineating the three age groups. This makes sense as the age groups were created so that there is roughly the same number of records in each group. But this does not mean that the boundaries are naturally discriminable boundaries. The COTS predictor is less susceptible to this fact as it is first predicts the age (in years), and we then bin this predicted age value into one of the three pre-defined groups. Thus, the COTS biographic predictor may include more discriminative information to train with and thus can ignore the arbitrary boundaries which may actually impede good prediction performance.

In summary, the following are the findings of the paper:

1) Label propagation is a viable method for imputing missing data in gallery records.
2) Suitably weighting individual attributes during label propagation stage is important.

We reiterate that the purpose of this work was to highlight the benefits of utilizing a graph-structure to model gallery records and *not* to improve state-of-the-art accuracy for biographic prediction.

VII. Summary

The primary purpose of this article is to motivate the use of graph-like structures to model the relationship between gallery records in a biometric database. Here, each gallery record is populated with both biometric (face) and biographic data (name, age-group, gender, ethnicity). While such a graph structure is likely to have several benefits, one specific benefit was explored in this work – the ability to impute missing biographic labels by exploiting both intra-record and inter-record information as characterized by the graph. As improvements to the label propagation algorithm is not the goal of our work, we adopted a label propagation scheme as-is from the literature to facilitate the prediction of missing data. The label propagation approach was observed to have success in this task as it was able to outperform a traditional face-image-based biographic predictor. This suggests the potential of the graph structure for use in identity-related tasks (biographic prediction, identity clustering, rapid recognition, *etc*.). In the future, we will develop sophisticated fusion methods to combine the label propagation scheme with a traditional face-image-based biographic predictor. This could improve the overall biographic prediction accuracy.

Acknowledgment

This project was supported by Award No. 2015-R2-CX-0005, from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice. Additionally, this work was supported in part by Michigan State University through computational resources provided by the Institute for Cyber-Enabled Research.

References

[1] A. K. Jain, A. A. Ross, and K. Nandakumar, *Introduction to Biometrics*. Springer Science & Business Media, 2011.

[2] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *International Conference on Automatic Face and Gesture Recognition*. IEEE, 2006, pp. 341–345.

[3] C. Otto, D. Wang, and A. Jain, "Clustering millions of faces by identity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[9] http://www.babynames1000.com/gender-neutral/

[4] T. Swearingen and A. Ross, "Predicting missing demographic information in biometric records using label propagation techniques," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2016, pp. 1–5.

[5] A. C. Gallagher and T. Chen, "Estimating age, gender, and identity using first name priors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2008, pp. 1–8.

[6] C. Shan, "Learning local binary patterns for gender classification on real-world face images," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 431–437, 2012.

[7] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 34–42, 2015.

[8] A. K. Jain, S. C. Dass, and K. Nandakumar, "Can soft biometric traits assist user recognition?" in *Defense and Security*, vol. 5404. International Society for Optics and Photonics, 2004, pp. 561–572.

[9] N. Almudhahka, M. Nixon, and J. Hare, "Human face identification via comparative soft biometrics," in *ISBA 2016 - IEEE International Conference on Identity, Security and Behavior Analysis*, 2016.

[10] N. Y. Almudhahka, M. S. Nixon, and J. S. Hare, "Unconstrained human identification using comparative facial soft biometrics," in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2016, pp. 1–6.

[11] D. Martinho-Corbishley, M. S. Nixon, and J. N. Carter, "Soft biometric retrieval to describe and identify surveillance images," in *ISBA 2016 - IEEE International Conference on Identity, Security and Behavior Analysis*, 2016.

[12] W.-T. Chu and C.-H. Chiu, "Predicting Occupation from Single Facial Images," in *IEEE International Symposium on Multimedia (ISM)*, 2014, pp. 9–12.

[13] H. Chen, A. C. Gallagher, and B. Girod, "What's in a name?" in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 6, 2013, pp. 3366–3373.

[14] A. Dantcheva, P. Elia, and A. Ross, "What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, pp. 441–467, 2016.

[15] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic Estimation from Face Images: Human vs. Machine Performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1148–1161, 2015.

[16] J.-C. Chen, A. Kumar, R. Ranjan, V. M. Patel, A. Alavi, and R. Chellappa, "A Cascaded Convolutional Neural Network for Age Estimation of Unconstrained Faces," in *International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2016, pp. 1–8.

[17] H. Ding, D. Huang, Y. Wang, and L. Chen, "Facial ethnicity classification based on boosted local texture and shape descriptions," *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pp. 1–6, 2013.

[18] G. Muhammad, M. Hussain, F. Alenezy, G. Bebis, A. M. Mirza, and H. Aboalsamh, "Race classification from face images using local descriptors," *International Journal on Artificial Intelligence Tools*, vol. 21, no. 05, p. 1250019, 2012.

[19] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar, "Describable visual attributes for face verification and image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1962–1977, 2011.

[20] A. Ambekar, C. Ward, J. Mohammed, S. Male, and S. Skiena, "Name-ethnicity classification from open sources," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 49–58, 2009.

[21] E. Makinen and R. Raisamo, "Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 541–547, 2008.

[22] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1955–1976, 2010.

[23] S. Fu, H. He, and Z.-G. Hou, "Race classification from Face: A Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 12, pp. 2483–2509, 2014.

[24] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1789–1801, 2012.

[25] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain, "Matching composite sketches to face photos: A component-based approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 191–204, 2013.

[26] A. K. Jain, K. Nandakumar, X. Lu, and U. Park, "Integrating faces, fingerprints, and soft biometric traits for user recognition," in *International Workshop on Biometric Authentication*, 2004, pp. 259–269.

[27] V. Tyagi and H. P. Karanam, "Fusing Biographical and Biometric Classifiers for Improved Person Identification," in *International Conference on Pattern Recognition (ICPR)*, 2012, pp. 2351–2354.

[28] H. S. Bhatt, R. Singh, and M. Vatsa, "Can combining demographics and biometrics improve de-duplication performance?" *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 188–193, 2013.

[29] P. S. Sudhish, A. K. Jain, and K. Cao, "Adaptive fusion of biometric and biographic information for identity de-duplication," *Pattern Recognition Letters*, vol. 84, pp. 199–207, 2016.

[30] A. A. Ross, K. Nandakumar, and A. A. Jain, *Handbook of Multibiometrics*, 2006, vol. 6.

[31] N. K. Ratha, J. H. Connell, and S. Pankanti, "Big Data approach to biometric-based identity analytics," *IBM Journal of Research and Development*, vol. 59, no. 2/3, pp. 4:1–4:11, 2015.

[32] "Think BIG Fusion of Biometric and Biographic Data In Large-Scale Identification Projects," WCC Smart Search & Match, Tech. Rep.

[33] W. Liu and D. Ruths, "What's in a Name? Using First Names as Features for Gender Inference in Twitter," *Analyzing Microtext: Papers from the 2013 AAAI Spring Symposium*, pp. 10–16, 2013.

[34] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Sch lkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 16, 2003, pp. 321–328.

[35] M. Rubinstein, C. Liu, and W. T. Freeman, "Annotation propagation in large image databases via dense image correspondence," in *European Conference on Computer Vision*, 2012, pp. 85–99.

[36] M. E. Houle, V. Oria, S. Satoh, and J. Sun, "Annotation propagation in image databases using similarity graphs," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 27, no. 1, pp. 288–311, 2013.

[37] J. Yu, X. Jin, J. Han, and J. Luo, "Collection-based sparse label propagation and its application on social group suggestion from photos," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 2, pp. 12:1–12:21, 2011.

[38] D. Liu, S. Yan, X. S. Hua, and H. J. Zhang, "Image retagging using collaborative tag propagation," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 702–712, 2011.

[39] J. Tang, M. Li, Z. Li, and C. Zhao, "Tag ranking based on salient region graph propagation," *Multimedia Systems*, vol. 21, no. 3, pp. 267–275, 2015.

[40] P. Mateos, "A review of name-based ethnicity classification methods and their potential in population studies," *Population, Space and Place*, vol. 13, no. 4, pp. 243–263, 2007.

[41] P. Skerry, *Counting on the census?: Race, group identity, and the evasion of politics*. Brookings Institution Press, 2000, vol. 56.

[42] J. Michael, "40,000 namen. anredebestimmung anhand des vornamens," *c't*, pp. 182–183, 2007.

[43] "Frequently occuring first names and surnames from the 1990 census," United States Census Bureau, Tech. Rep., 1995.

[44] D. L. Word, C. D. Coleman, R. Nunziata, and R. Kominski, "Demographic Aspects of Surnames From Census 2000," United States Census Bureau, Tech. Rep.

[45] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[46] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.