18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

A Bayesian Idealization of Entity Resolution

James P. Ferry
Metron, Inc.
Reston, Virginia 20190
Email: ferry@metsci.com

Darren Lo
Metron, Inc.
Reston, Virginia 20190
Email: lo@metsci.com

Thomas Seaquist
Metron, Inc.
Reston, Virginia 20190
Email: seaquist@metsci.com

***Abstract*—Network theory has progressed a long way since the Erdős–Rényi model, identifying many important real-world phenomena that a good random graph model should capture, and producing more realistic models to capture many of them. However, these models are largely limited to the domain of simple networks—nodes and links only—leaving remaining complications outside the realm of theory. In such cases, a practitioner with complicated data is left to make decisions or apply algorithms to compensate for these issues without the benefit of an underlying model. In this paper, we develop a simple generative model of the entity resolution problem. Noting its similarity to the association problem in data fusion, we develop principled inference equations for entity resolution analogous to those developed for data association. The framework for this effort is a ground-truth model for object states and for the network which links them, together with a Dirichlet process model for how the observed aliases of the objects are distributed among the observed transactions between them. The paper focuses on the derivation of the inference equations, and the result is demonstrated on an illustrative example. Because the framework is based on rigorous probabilistic models, it is particularly well suited to ambiguous scenarios in which no single entity resolution hypothesis is stands out as the correct one.**

I. Introduction

A critical issue complicating the analysis of collections of data is identifying references that belong to the same underlying entity. It is the fundamental problem that arises when joining two databases that refer to overlapping sets of entities. Even within a single database, good entity resolution is a pre-requisite for any subsequent analysis that involves knowing which data is associated with which person, device, corporation, or vehicle. For example, in a bibliographic database containing references to published papers and their authors, an analysis may begin by considering various people's complete works. This can be challenging. Is John Smith on publication a the same Jonathan Smith on publication b? Does the "John Smith" of publication c refer to the same author as the "John Smith" of publication a? To which entity does the abbreviated J. Smith on publication d refer?

The partitioning of a collection of entity references into sets corresponding to the same ground-truth entity is termed *entity resolution*. Ironically, the term "entity resolution" itself suffers from the very ailment it seeks to correct. In different communities it known by various aliases—record linkage, duplicate detection, deduplication, entity disambiguation/linking, etc.

Older approaches to entity resolution employ pairwise attribute comparisons between references, and are limited to identifying typographical or parse errors. These approaches may be considered to lie within the realm of *clustering*. We will limit the term "entity resolution" to refer only to more sophisticated approaches that exploit both attributes and some kind of network structure: e.g., both authors' published names and the co-authorship network among them. For example, a shipment database containing all shipments along with the corresponding shipper and consignee forms a graph where the nodes are the trade entity references and the edges represent observed shipments.

The *association* problem that arises in the data fusion community is similar to the entity resolution problem, and the purpose of this paper is to develop an entity-resolution analog of the principled data-fusion association methodology. Data association is the resolution of which observations on different sensors refer to the same object. Clustering approaches do not apply because they do not enforce structural constraints, such as no split or merged measurements. This problem is over 50 years old [1], and was originally approached with an *algorithmic* mindset, just as entity resolution is today. However, in a series of papers beginning with [2] in 1990, and reaching maturity in the 2000s (e.g., [3], [4]), Mori, Chong, and others superseded the algorithmic mindset with a *probabilistic* one. Rather than computing an association that maximizes an *ad hoc* score, they defined a model in which the probability of an association hypothesis, given the observed data, is well defined. In this framework, an association is considered optimal not (a) because it is what a well-regarded algorithm produces, or (b) because it maximizes some "score," but (c) because it has the MAP (Maximal *A posteriori* Probability) according to a very natural model.

The MAP algorithm itself is nearly identical to the score-based approach. The only difference is that MAP prescribes an "adaptive threshold" for accepting association hypotheses. This was demonstrated to improve performance [5], but from an algorithmic perspective, it amounted only to a helpful tweak. The more profound impact of the probabilistic framework was that it generalized to non-kinematic data types. The XMAP (eXtended MAP) method [6] describes the statistical information required to incorporate non-kinematic features into association models and algorithms [4], [7].

There are a number of effective entity resolution algorithms. In [8] an agglomerative clustering algorithm that employs relational similarity between nodes updates the graph upon each agglomeration. While this technique is a good one, it does not necessarily arrive at a globally optimal solution since

no score is ever assigned to a particular resolution hypothesis. In an attempt to assign such a score to a hypothesis, undirected graphical models (Markov random fields) have been used to assign a probability to any given hypothesis. In [9] Markov logic, a combination of first-order logic and random fields, is used to assign probabilities to hypotheses. In [10], [11] a discriminative model uses conditional random fields to produce such probabilities. While these models assign probabilities to entity resolution hypotheses, because the models are not generative, one cannot interpret the probabilities they produce as belonging to events whose causes are understood. Instead, we seek a generative model for the scenarios that require entity resolution, so that we may scientifically study the causal mechanisms and inference required for entity resolution.

Modeling entity resolution scenarios is more complex than data association because of the network component of the problem. The simplest probability distribution over networks is the Erdős–Rényi model $\mathcal{G}(n,p)$ [12]. A draw from $\mathcal{G}(n,p)$ is a graph with n nodes, with edges instantiated independently with probability p. This simple model for networks exhibits the small-world property of real-world networks [13], but lacks other realistic attributes such as highly unequal degree distribution and large clustering coefficient [14]. Nevertheless, the Erdős–Rényi model is the natural network model to begin with because it is simple and provides a baseline for future development. We combine this network model with a Dirichlet process for entity fragmentation (described below) and a Gaussian spatial model to produce a relatively tractable scenario for studying entity resolution.

A. Model specification

The idea of our model is as follows. We imagine that there is an unobserved set of *objects* and that certain pairs of objects have a relationship which permits some number of *transactions* to be generated. For each transaction a pair of *aliases* records the objects involved in the transaction, but only imperfectly—the same object may have multiple aliases. We restrict our attention to the case where no two objects may share the same alias, however.

We let $\mathcal{X}$ denote the state space for objects, and $\mathcal{Z}$ denote the state space for their aliases. These may be the same space (for example, a space of strings) or different ones ($\mathcal{X}$ could be a rich ground-truth space, whereas $\mathcal{Z}$ could be the limited space in which we make observations). If there are n objects, we label them $j = 1$ through n, using $[n] = \{1, 2, \ldots, n\}$ to denote the set of objects. The state of object j is denoted x_j. Relations among the objects are represented by a graph G with n nodes, and its edge set E are the pairs of objects which are permitted to have transactions. We model G with the Erdős–Rényi process $G \sim \mathcal{G}(n,p)$. If a pair of objects is connected by an edge $e \in E$, then that pair generates a Poisson number of transactions, $k_e \sim \mathrm{Po}(\lambda)$.

To motivate the mechanism that produces aliases from objects, we stipulate a few desiderata. First is an exchangeability criterion: for a fixed object state x, the probability of obtaining a particular multiset of alias states should be independent of the order of transactions. Second, alias states have a nonzero probability of being generated repeatedly. Given the prevalence of power-law scaling in real-world data, which can be generated by linear preferential attachment, it seems appropriate to postulate a similar effect for aliases as well: when considering transactions in any ordering, the probability of observing an alias state again should be roughly proportional to the number of times it has already been seen. This suggests that we model the generation of aliases as a Blackwell–MacQueen urn scheme [15], or equivalently using a Dirichlet process, independently for each object. (Note that the Dirichlet process itself does not produce power-law behavior; we have merely drawn inspiration from linear preferential attachment schemes of Yule type [16]. A possible extension of this work would be to use the two-parameter Pitman–Yor process, which is a generalization of the Dirichlet process that does produce power-law behavior [17], [18].)

Specifically, we assume that there exists a map $x \mapsto H_x^0$ taking object states to probability measures on $\mathcal{Z}$. Then independently for each object $j \in [n]$, we draw a random probability measure $H_j \sim \mathrm{DP}(H_{x_j}^0, \theta)$, which is (almost surely) discrete. The aliases for x are then generated as i.i.d. draws from H_x. To avoid the complication of different objects generating the same alias state, we stipulate that the base measures H_x^0 have no atoms.

If we let $P_{\mathcal{X}}$ denote a prior distribution on the state space $\mathcal{X}$, then the full model specification is as follows:

$$G \sim \mathcal{G}(n,p) \tag{1a}$$
$$k_e \sim \mathrm{Po}(\lambda) \qquad e \in E \tag{1b}$$
$$x_j \sim P_{\mathcal{X}} \qquad j \in [n] \tag{1c}$$
$$H_{x_j} \sim \mathrm{DP}(H_{x_j}^0, \theta) \qquad j \in [n] \tag{1d}$$
$$z_{j,e,\epsilon} \sim H_{x_j} \qquad j \in e \in E; \epsilon \in [k_e] \tag{1e}$$
$$T_Z^+ = \{\{z_{j,e,\epsilon}, z_{j',e,\epsilon}\} : \{j, j'\} = e \in E, \epsilon \in [k_e]\} \tag{1f}$$

In (1f), the observed data T_Z^+ is a multiset of transaction records. A number of parameters occur in the model: n, p, λ, θ, and any that occur within the prior $P_{\mathcal{X}}$ and the family of base measures H_x^0. We keep some of these parameters, and integrate others against corresponding prior distributions.

For $j \in [n]$, let $\mathbf{z}_j = \{z_{j,e,\epsilon} : j \in e \in E, \epsilon \in [k_e]\}$ be the set of alias states generated by j. Then the $\mathbf{z}_j$ are (almost surely) pairwise disjoint and thus the collection of nonempty $\mathbf{z}_j$ is a partition of the set of alias states that occur within T_Z^+. The inference problem is to determine the probability of such a partition from the observed data T_Z^+. Section II derives this probability. An illustrative example is shown in Section III and directions for future work are discussed in IV.

II. Derivation

We let A be an *association* of the observed aliases—a specification of which aliases arise from the same object. We let Z denote the observed data: this will end up being equivalent to T_Z^+ in (1f), but rewritten in a form that facilitates the derivation of the inference equations. We wish to derive the probability of

A given Z: i.e., $\Pr(A|Z)$. This conditional probability is given by $\Pr(A|Z) = \Pr(A,Z)/\Pr(Z)$, where the factor $1/\Pr(Z)$ is just a normalization constant. To calculate $\Pr(A,Z)$ we formulate an observation model $\Pr(A,Z|X)$ in terms of a hidden, ground-truth state X, as well as a prior distribution $\Pr(X)$ on X. Integrating out X yields

$$\Pr(A|Z) \propto \int \Pr(A,Z|X)\Pr(X)\,dX. \tag{2}$$

The derivation of $\Pr(A|Z)$ is analogous to the derivation of the traditional data fusion association probability as presented in [7], [19].

A. Association probability

We will define the state X in (2) as $X = (E^+, \mathbf{x}, n, \vec{p})$, where $\vec{p}$ is a vector those model parameters which we wish to integrate out of the problem, n the number of objects, $\mathbf{x}$ is an array of object states x_j for $j \in [n]$, and E^+ is the multiset of edges of a multigraph with nodes $[n]$. We let E be the edge set (without multiplicity) corresponding to E^+. Then each edge $e \in E$ may be represented as $e = \{j, j'\}$, and has a multiplicity k_e. We may represent each $e^+ \in E^+$ as $e^+ = (e, \epsilon)$ where $\epsilon \in [k_e]$. We will assume E^+ and $\mathbf{x}$ are conditionally independent given n and $\vec{p}$. We may think of this as the structure of trade networks being independent of the names of the companies involved. Furthermore, we assume that the components x_j of $\mathbf{x}$ are i.i.d. random variables that do not depend on $\vec{p}$. Then

$$\Pr(X) = \Pr(E^+|n,\vec{p})\Pr(n,\vec{p})\prod_{j=1}^{n}\Pr(x_j). \tag{3}$$

The data Z in (2) will have the form $Z = (T^+, \mathbf{z})$, where $\mathbf{z}$ is an array (of length m) of alias states z_i, and T^+ is a multiset of transactions. This separates the structural information in T_Z^+ from the state information. We let T be the transaction set (without multiplicity) corresponding to T^+. Then each transaction $t \in T$ may be represented as $t = \{i, i'\}$, and has a multiplicity w_t. We may represent each $t^+ \in T^+$ as $t^+ = (t, \tau)$ where $\tau \in [w_t]$. Each alias i is a corrupted version of some object j: we define $a(i) \to j$ to be the assignment function that maps aliases to their objects. This assignment function need not be surjective—we may well have $a^{-1}(j) = \varnothing$ for many objects $j \in [n]$. Given $j = a(i)$ and $j' = a(i')$ then each transaction $t^+ = (\{i, i'\}, \tau) \in T^+$ corresponds to some edge $e^+ = (\{j, j'\}, \epsilon) \in E^+$. Finally, the association A is simply a partition of the m aliases, which is an unlabeled version of the assignment function a.

To derive the conditional probability $\Pr(A, Z|X)$ required in (2), we begin with letting $E_j^+ \subseteq E^+$ be the multiset of edges containing the node corresponding to object j and define

$$d_j = |E_j^+| = \sum_{e \in E_j} k_e$$

to be the degree of node j. (Note: we will be overloading "d" to refer to various multigraph degrees, trusting the meaning to be clear from context.) We partition the d_j edges at node j into blocks (which will later be identified with aliases). The edge partition induced by the DP-distributed measure H_{x_j} is distributed according to the Chinese Restaurant Process (CRP) or Ewens distribution. The CRP gives the probability of a partition of d_j labeled objects into a set B of unlabeled blocks b:

$$\Pr(B|d_j) = \frac{\theta^{|B|}\Gamma(\theta)}{\Gamma(\theta + d_j)}\prod_{b \in B}\Gamma(|b|). \tag{4}$$

Let $m_j = |B|$ denote the number of blocks for node j, and suppose we label (or order) the blocks 1 through m_j. We let $d_{jr} = |b|$ denote the number of edges in block r for $r = 1$ to m_j, with

$$\sum_{r=1}^{m_j} d_{jr} = d_j.$$

Whereas (4) gives the probability of an (unordered) set of blocks, we let B_j denote the (ordered) array of blocks at node j. The probability of any B_j is $1/m_j!$ times the probability of B because each of the $m_j!$ orderings is equally likely.

Let $z_{jr} \sim H_{x_j}$ denote the state assigned to block r, and $\Pr(z_{jr}|x_j)$ denote the probability density of z_{jr} given the state x_j of object j. Then the joint probability density of B_j and the locations $\mathbf{z}_j$ of all the blocks arising from node j is

$$\Pr(B_j, \mathbf{z}_j|d_j, x_j) = \frac{\theta^{m_j}\Gamma(\theta)}{\Gamma(\theta + d_j)\,m_j!}\prod_{r=1}^{m_j}\Gamma(d_{jr})\Pr(z_{jr}|x_j).$$

The product of this over all nodes $j \in [n]$ gives the joint probability density of all the block arrays $\vec{B}$ and all the alias states $\vec{\mathbf{z}}$:

$$\Pr(\vec{B},\vec{\mathbf{z}}|E^+,\mathbf{x}) = \theta^m\prod_{j=1}^{n}\left(\frac{\Gamma(\theta)}{\Gamma(\theta+d_j)\,m_j!}\prod_{r=1}^{m_j}\Gamma(d_{jr})\Pr(z_{jr}|x_j)\right). \tag{5}$$

We now translate $\vec{B}$ in (5) in terms of T^+ and a. There are $m!/(m_1!m_2!...m_n!)$ ways to assign global indices 1 through m to the blocks in $\vec{B}$ while preserving the local ordering for each j. We may also take equivalence classes under edge labeling. There are $k_e!$ ways of labeling the k_e edges for each $e \in E$. Combining all these into a single, unlabeled class would overcount by the product of $w_t!$ over $t \in T$. Rather than dividing out by this, however, we retain the local labeling $\tau = 1$ to w_t for each $t \in T$. We may now rewrite $\vec{\mathbf{z}}$ as $\mathbf{z}$, which is indexed from $i = 1$ to m. Thus we replace z_{jr} by z_i. We also replace d_{jr} by d_i, the degree in T^+ of node i, trusting the notation to keep this from being confused with d_j. The result is

$$\Pr(T^+, a, \mathbf{z}|E^+,\mathbf{x}) = \frac{\theta^m}{m!}\prod_{e\in E}k_e!\prod_{i=1}^{m}\Gamma(d_i)\times \prod_{j=1}^{n}\left(\frac{\Gamma(\theta)}{\Gamma(\theta+d_j)}\prod_{i\in a^{-1}(j)}\Pr(z_i|x_j)\right). \tag{6}$$

Here we note that E^+ is determined by T^+ and a. Thus (6) holds only for values that are consistent. Otherwise, the probability density is zero.

B. Integration

We have suppressed the dependence on n and $\vec{p}$ in the derivation of (6), but we restore it now in order to integrate (6) against the prior (3):

$$\begin{aligned}\Pr(T^+,a,\mathbf{z}) = \iint \Pr(T^+,a,\mathbf{z}|E^+,\mathbf{x},n,\vec{p})\times \\ \Pr(E^+|n,\vec{p})\Pr(n,\vec{p})\prod_{j=1}^{n}\Pr(x_j)\,d\mathbf{x}d\vec{p}.\end{aligned} \tag{7}$$

Note that it is unnecessary to sum over n or E^+: the only value of n that contributes is the one determined by a, and the only value of E^+ that contributes is the one determined by T^+ and a, so we interpret n and E^+ to mean these values in (7). In particular, note that $|E^+| = |T^+|$. To integrate $\mathbf{x}$ out of (7) we let

$$\Pr(\mathbf{z}_\alpha) = \int \prod_{i\in\alpha}\Pr(z_i|x)\Pr(x)\,dx \tag{8}$$

denote the probability that a collection of locations $\mathbf{z}_\alpha$ arises from a common state. Then

$$\begin{aligned}\Pr(T^+,a,\mathbf{z}) = \int \frac{\theta^m}{m!}\prod_{e\in E}k_e!\prod_{i=1}^{m}\Gamma(d_i)\times \\ \prod_{j=1}^{n}\frac{\Gamma(\theta)\Pr(\mathbf{z}_{a^{-1}(j)})}{\Gamma(\theta+d_j)}\Pr(E^+|n,\vec{p})\Pr(n,\vec{p})\,d\vec{p}.\end{aligned}$$

The model for E^+ is parameterized by n, p and λ. First, we draw a "trading partner" graph G from the Erdős–Rényi process $\mathcal{G}(n,p)$. Then for each edge of G, we generate a Poisson number of edges, $\mathrm{Po}(\lambda)$, in E^+. Thus E comprises the edges of G, and we define E_0 to be the those edges for which the number of transactions $\mathrm{Po}(\lambda)$ is positive.

$$\Pr(E^+|n,p,\lambda) = \left(pe^{-\lambda}\right)^{|E_0|}q^{\binom{n}{2}-|E_0|}\lambda^{|T^+|}\Big/\prod_{e\in E}k_e!,$$

where $q = 1-(1-e^{-\lambda})p$.

We also collect functions a with the same range into equivalence classes, denoting the resulting association A, which is a partition of the integers $i\in[m]$. The values of A and T^+ determine the multigraph E^+ modulo its node labeling. We let n_0 denote the number of nodes in the multigraph E^+ determined by A and T^+. The association A encompasses assignment functions a with any number of nodes $n\geq n_0$. In particular, there are $n!/(n-n_0)!$ ways to relabel the n_0 nodes of E^+ using the labels $j=1$ to n. Therefore

$$\begin{aligned}\Pr(T^+,A,\mathbf{z}) = \prod_{i=1}^{m}\Gamma(d_i)\int\sum_{n=n_0}^{\infty}\frac{\theta^m}{m!}\left(pe^{-\lambda}\right)^{|E_0|}\times \\ \lambda^{|T^+|}\prod_{\alpha\in A}\frac{\Gamma(\theta)\Pr(\mathbf{z}_\alpha)}{\Gamma(\theta+d_\alpha)}\frac{n!\Pr(n,\vec{p})}{(n-n_0)!}q^{\binom{n}{2}-|E_0|}\,d\vec{p}.\end{aligned} \tag{9}$$

We use the beta prior $\Pr(p|n) = \mathrm{Beta}(p;\delta,n)$, and log-uniform priors, $\Pr(\lambda)\propto 1/\lambda$, and $\Pr(n)\propto 1/n$. The (asymptotic) mean degree δ remains as a parameter, as does θ. With these priors we can integrate out the model parameters $\vec{p}$ in (9) to find that

$$\Pr(A|T^+,\mathbf{z}) \propto \left(\sum_{n=n_0}^{\infty}F(n)\right)\prod_{\alpha\in A}\frac{\Gamma(\theta)\Pr(\mathbf{z}_\alpha)}{\Gamma(\theta+d_\alpha)}, \tag{10}$$

where

$$\begin{aligned}F(n) = \frac{(n-1)!}{(n-n_0)!}\frac{1}{\mathrm{B}(\delta,n)}\int_0^\infty\int_0^1 p^{\delta-1}(1-p)^{n-1} \\ \left(pe^{-\lambda}\right)^{|E_0|}\left(1-(1-e^{-\lambda})p\right)^{\binom{n}{2}-|E_0|}\lambda^{|T^+|-1}\,dpd\lambda.\end{aligned} \tag{11}$$

Let $F^*(n)$ be an approximation to $F(n)$ in which the $(1-e^\lambda)$ factor in the integrand is replaced by 1. We may interpret the version of (10) with $F(n)$ replaced by $F^*(n)$ as the probability of both the association A and the event that all the edges of G yielded at least one transaction, which is a lower bound on $\Pr(A|T^+,\mathbf{z})$. Likewise, $F^*(n)\leq F(n)$. The double integral in (11) splits into two easily evaluated single integrals. Thus

$$F^*(n) = \frac{\Gamma(|T^+|)}{|E_0|^{|T^+|}}\varphi(n),$$

where

$$\varphi(n) = \frac{(n-1)!}{(n-n_0)!}\frac{\mathrm{B}\big(\delta+|E_0|,n(n+1)/2-|E_0|\big)}{\mathrm{B}(\delta,n)}.$$

To estimate the error in the approximation $F(n)\approx F^*(n)$, we note that the peak of the integrand of $F^*(n)$ occurs at $\lambda_c = (|T^+|-1)/|E_0|$ and $p_c = 2(|E_0|+\delta-1)/(n(n+1)+2\delta-4)$. Substituting these values into the ratio of the integrands of $F(n)$ to $F^*(n)$ yields the factor

$$\rho(n) = \left(1+\frac{(|E_0|+\delta-1)e^{(1-|T^+|)/|E_0|}}{n(n+1)/2-|E_0|-1}\right)^{\binom{n}{2}-|E_0|}, \tag{12}$$

which is roughly $\exp(|E_0|e^{-|T^+|/|E_0|})$. This implies that the approximation $F(n)\approx F^*(n)$ works well in the regime $|E_0|\log|E_0| \lesssim |T^+|$. In particular, it works better for hypotheses A with small associated values of $|E_0|$.

When using the approximation $F(n)\approx F^*(n)$, the sum over n in (10) requires a sum over $\varphi(n)$. We let

$$\Phi = \sum_{n=n_0}^{\infty}\varphi(n).$$

For a fixed value of δ, the sum Φ involves only the values of n_0 and $|E_0|$ for a given association hypothesis A. As such, the values of Φ may be computed numerically and stored as needed without incurring undue memory overhead. In the course of computing Φ for specific values of n_0 and $|E_0|$ we may also store

$$n_c = \operatorname*{argmax}_n \varphi(n).$$

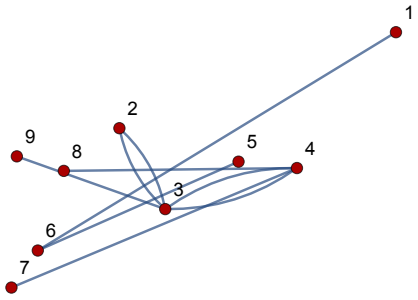

Fig. 1. Nine aliases are observed at the plotted locations, as are transactions between them. Which observations arise from the same ground-truth objects?

The values of n near n_c are the ones that dominate, so we may use the value $n = n_c$ in (12) to compute a correction factor for making the approximation $F(n) \approx F^*(n)$. The final result is the following simple formula for the posterior probability of A:

$$\Pr(A|T^+, \mathbf{z}) \propto \frac{\Phi\, \rho(n_c)}{|E_0|^{|T^+|}} \prod_{\alpha \in A} \frac{\Gamma(\theta)\Pr(\mathbf{z}_\alpha)}{\Gamma(\theta + d_\alpha)}. \qquad (13)$$

C. Spatial component

To evaluate the value of (8) required in (13) we let $\mathcal{X} = \mathcal{Z} = \mathbb{R}^r$ for some fixed r. We let the observation error $\Pr(z|x)$ be an isotropic Gaussian with standard deviation σ:

$$\Pr(z|x) = \mathcal{N}(z; x, \sigma^2 I) = (2\pi\sigma^2)^{-r/2} e^{-|z-x|^2/(2\sigma^2)}.$$

For the prior there are two natural choices: Gaussian and uniform.

If we use a unit Gaussian prior $\Pr(x) = \mathcal{N}(x; 0, I)$ then

$$\Pr(\mathbf{z}_\alpha) = (2\pi w_\alpha)^{r/2} \mathcal{N}(\mu_\alpha; 0, I) \prod_{i \in \alpha} \mathcal{N}(z_i; \mu_\alpha, \sigma^2 I), \quad (14)$$

where

$$w_\alpha = \frac{\sigma^2}{|\alpha| + \sigma^2} \quad \text{and} \quad \mu_\alpha = \frac{|\alpha|\bar{z}_\alpha + \sigma^2}{|\alpha| + \sigma^2},$$

with $\bar{z}_\alpha$ denoting the mean of z_i over $i \in \alpha$.

On the other hand, suppose we let $\Pr(x) = I_R(x)/V$: i.e., a function that equals $1/V$ in a region R of volume V and is zero elsewhere. This is the prior used in [3] and elsewhere which produces results similar to the previous *ad hoc* association algorithms. Provided σ is sufficiently small and the z_i are located away from the boundary of R, we may approximate the integral over R with an unbounded integral. To approximate a unit Gaussian, we let $V = (4\pi)^{r/2}$ (cf. [19]). Then

$$\Pr(\mathbf{z}_\alpha) = (\sigma^2/(2|\alpha|))^{r/2} \prod_{i \in \alpha} \mathcal{N}(z_i; \bar{z}_\alpha, \sigma^2 I). \qquad (15)$$

The advantage of (15) over (14) is that it is simpler and does not distinguish 0 as a special location. Although (15) is only valid when $\sigma \lesssim 1$, this is the region of interest. However, to assess the behavior near the limit where location data carries no information it is necessary to use (14).

TABLE I
TEN MOST PROBABLE ASSOCIATIONS. ✓ = TRUTH.

	$\Pr(A)$	A
	0.250	{{1}, {2, 9}, {3}, {4, 5}, {6, 7, 8}}
	0.111	{{1}, {2, 9}, {3, 8}, {4, 5}, {6, 7}}
	0.072	{{1}, {2, 8, 9}, {3}, {4, 5}, {6, 7}}
	0.059	{{1}, {2, 9}, {3}, {4, 5}, {6, 7}, {8}}
	0.048	{{1, 4, 5}, {2, 9}, {3}, {6, 7, 8}}
✓	0.048	{{1}, {2}, {3}, {4, 5}, {6, 7, 8}, {9}}
	0.034	{{1}, {2}, {3}, {4, 5}, {6, 7}, {8, 9}}
	0.024	{{1, 4}, {2, 9}, {3}, {5}, {6, 7, 8}}
	0.023	{{1}, {2}, {3}, {4, 5}, {6, 7, 8, 9}}
	0.021	{{1, 4, 5}, {2, 9}, {3, 8}, {6, 7}}

III. Example

The formula (13) does not immediately yield a practical algorithm for entity resolution. We may compute association probabilities, but there are an exponentially large number of associations to evaluate. This problem arises in the association algorithms used in the data fusion community as well, and computing the MAP probability exactly for more than two sensors is known to be NP-hard, so approximate methods are required [20]. In the two-sensor case there are efficient integer programming algorithms to find optimal solutions [21], but there is no analog of such a case in the entity resolution problem. We will not describe a practical algorithm here, but only provide an illustrative example of the method on a small case.

Suppose we are given the data T_Z^+ shown in Figure 1. Here there are nine aliases with states that are known to be noisy, and only a few transactions observed between them. In this scenario there is not enough data to be confident about any one entity resolution hypothesis. An analyst would be able to rule out pairing 1 with 7 (too far apart), but then would have to resort to some combination of running an entity resolution algorithm and relying on *ad hoc* reasoning. For example, "1 and 5 are pretty far apart, but are each connected only to 6, so they might paired," or "6 and 7 are close, but have no common neighbor... the same applies to 8 and 9—each of these might be paired too." However, it is difficult to sort out the delicate interplay between network and state information. For example, using the parameters discussed below we find that the probability of 6 and 7 arising from the same ground-truth object is 93%, but the probability for 8 and 9 is only 21%.

We use the parameters $\sigma = 0.5$ and $\theta = 1$ in the association probability formula (13). Of the $B_9 = 21,147$ possible partitions A of nine elements, 5017 are compatible with the multigraph T^+ depicted in Figure 1 (having no edges within the clusters $\alpha \in A$). We compute $\Pr(A|T^+, \mathbf{z})$ using (13) for each of these compatible association hypotheses A. The ten most probable hypotheses are listed in Table I.

The most probable association has a probability more that twice that of any other. However, its probability is still only 25%, and the correct association is more likely be something

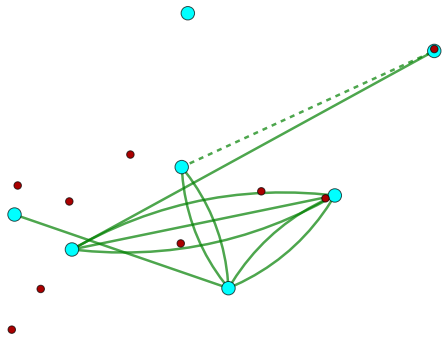

Fig. 2. Seven ground-truth objects and their graph G (shown with transaction multiplicities E^+). The aliases and transactions that arise from the objects and edges can be seen by comparing with Fig. 1 and Table I.

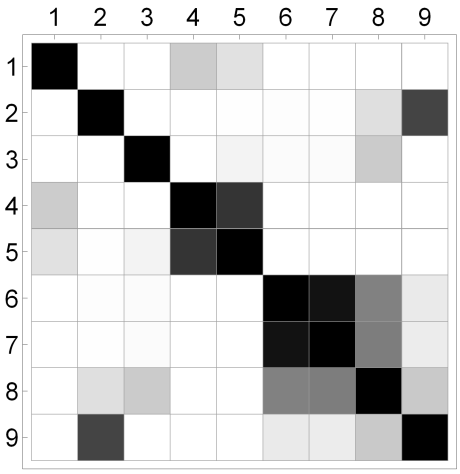

Fig. 3. Pair probabilities

TABLE II
TEN MOST PROBABLE PAIRS. ✓= TRUTH.

	$\Pr(\{i,i'\})$	$\{i,i'\}$
✓	0.927	{6, 7}
✓	0.795	{4, 5}
	0.733	{2, 9}
✓	0.509	{7, 8}
✓	0.494	{6, 8}
	0.211	{8, 9}
	0.203	{3, 8}
	0.202	{1, 4}
	0.125	{2, 8}
	0.117	{1, 5}

else. In fact, the data was simulated using the generative model we have defined here, so there is a correct association (indicated by the ✓), and it is not, in fact, not the most probable one in Table I, but the sixth most probable. Figure 2 shows the ground-truth scenario corresponding to Figure 1. There were $n = 7$ objects, although only $n_0 = 6$ of them were involved in transactions. The dotted edge e was part of the graph $G \sim \mathcal{G}(7, 0.3)$, but $k_e = 0$ transactions were drawn from $\mathrm{Po}(2.5)$ for this edge, so it does not belong to E.

For each pair $\{i, i'\}$ of aliases we may sum $\Pr(A|T^+, \mathbf{z})$ over all A that assign i and i' to the same cluster $\alpha \in A$. The results are depicted in Figure 3 with white denoting 0 probability and black denoting 1. The ten most probable pairs $\{i, i'\}$ of aliases are listed in Table II. Note that there are only four true pairs (indicated by ✓), and that they occur in the top five most likely. The other pair, $\{2, 9\}$ is the anomaly that occurs in all five associations more probable than the correct one in Table I. Referring to Figure 2, we see that the objects that produced aliases 2 and 9 are rather well separated, but happened to generate aliases in the direction of the other object. These objects also both happened to be connected to the same object, and to no others, and that object happened not to split into multiple aliases. These coincidences produced the illusion of 2 and 9 arising from a common object. The benefit of having a model to use with these small-data cases is that it can incorporate the possibility of all such coincidences automatically and assign meaningful probabilities.

For comparison, Figure 4 gives the output of the Bhattacharya–Getoor algorithm [8] for different values of a parameter α that controls the relative weight of the structure and state data. For example, with $\alpha = 0.1$ the state data is given high weight, and aliases 8 and 9 (being so close together in Figure 1) are tightly bound in the dendrogram; whereas for larger α the structure information (i.e., that aliases 8 and 9 have no common neighbors) increasingly overules the state information and puts aliases 8 and 9 in different branches of the dendrogram.

Each dendrogram provides a range of possible associations, depending on how coarse a solution the user desires. For the $\alpha = 0.2$ dendrogram there is a fairly large range (from 0.2972 to 0.3717 in the coarseness parameter), where the result is the same as the most probable association in Table I. This dendrogram correctly allows for aliases 6, 7, and 8 to be in the same cluster—6 and 7 due to their spatial proximity; 7 and 8 due their common neighbor. On the other hand, the $\alpha = 0.1$ dendrogram, which emphasizes space, binds 6 and 7 tightly, but not 7 and 8, whereas the opposite holds for the $\alpha = 0.3$ case.

Only 9 of the 5017 possible associations are represented in a given dendrogram in Figure 4. Even in the best case ($\alpha = 0.2$), the correct association is not included because 2 and 9 bind a little more tightly than 4 and 5. Of course, a practical algorithm like this cannot afford to sum over all possible associations. One benefit of our generative methodology is that it serves as a baseline for studying entity resolution and assessing algorithm performance. Thus we may compare how tightly the various pairs bind in Figure 4 with the probabilities in Table II.

IV. Conclusions and Future Directions

The model presented in this paper would be directly useful for cases where there are a very small number of aliases,

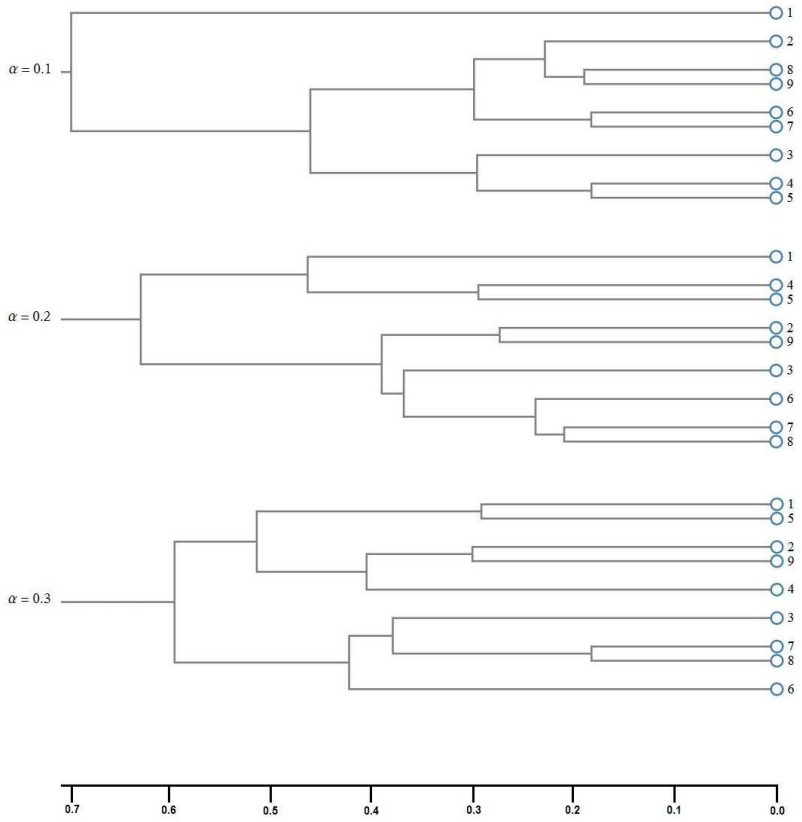

Fig. 4. Dendrograms from the Bhattacharya–Getoor algorithm

provided the object and alias spaces are accurately represented. When combined with appropriate methods of hypothesis management it could be useful for much larger cases, leading to algorithms which not only produce a single, best entity resolution hypothesis, but also provide probabilities for various events of interest, such as two specific aliases referring to the same object.

The larger purpose of this paper, however, is to establish a simple, standard model of an entity resolution scenario so that the phenomenon itself may be better understood. For example, we employ Gaussian distributions not because they are realistic representations of the type of data that usually requires entity resolution (such as names), but because Gaussian distributions in $\mathbb{R}^r$ provide the simplest, most generic encapsulation of an uncertain spatial measurement. Thus, we think of the model as a *Bayesian Idealization* of Entity Resolution.

Future work could discuss the phenomenology of entity resolution, as expressed in this model, in a manner similar to the study of the Erdős–Rényi model $\mathcal{G}(n,p)$ itself, establishing results analogous to, say, the emergence of the giant component [12]. Another important line of inquiry would establish limits on entity resolution performance, as expressed, for example, by results about the expected entropy of the posterior distribution over A.

The Erdős–Rényi model can be used as a network model, but has many shortcomings [22]. A number of more realistic models have been studied, and we expect entity resolution modeling to evolve in a similar fashion. Indeed, one way for it to evolve would be to use these more realistic network models. For example, the Barabási–Albert model has more realistic degree distribution, in which some nodes have very large degree [23]. This is an important feature to represent in entity resolution, because a common high-degree neighbor is not very informative: e.g., two aliases are not much more likely to refer to the same person just because both received shipments from Amazon.

Bayesian idealization is useful in other network problems as well. There are good algorithms for link detection (e.g., [24]), but a lack of formal observation models corresponding to the plethora of network models that could be used to represent ground truth. On the other hand, the Stochastic Block Model (SBM) is simple idealization of the community detection problem [25], [26], [27]. Recent work with the SBM has demonstrated phase transitions and the limits of inference

in community detection [28]. It is our hope that as various problems are idealized in models like these and the one in this paper, they will further each other's development to the mutual benefit of all.

REFERENCES

[1] R. W. Sittler, "An optimal data association problem in surveillance theory," *IEEE Transactions on Military Electronics*, pp. 125–139, Apr. 1964.

[2] C.-Y. Chong, S. Mori, and K.-C. Chang, "Distributed Multitarget Multisensor Tracking," in *Multitarget–Multisensor Tracking: Advanced Applications*, Y. Bar-Shalom, Ed. Boston: Artech, 1990, ch. 8.

[3] S. Mori and C.-Y. Chong, "Track-To-Track Association Metric—I.I.D.-Non-Poisson Cases—," in 6^{th} *International Conference on Information Fusion*, Jul. 2003.

[4] C.-Y. Chong and S. Mori, "Metrics for Feature-Aided Track Association," in 9^{th} *International Conference on Information Fusion*, Jul. 2006.

[5] L. Stone, T. M. Tran, and M. L. Williams, "Improvement in track-to-track association from using an adaptive threshold," in 12^{th} *International Conference on Information Fusion*, Jul. 2009.

[6] L. D. Stone, M. L. Williams, and T. M. Tran, "Track-to-Track Association and Bias Removal," in *SPIE AeroSense International Conference*, Apr. 2002.

[7] J. Ferry, "XMAP: Track-to-Track Association with Metric, Feature, and Target-type Data," in 9^{th} *International Conference on Information Fusion*, Jul. 2006.

[8] I. Bhattacharya and L. Getoor, "Collective entity resolution in relational data," *Transactions on Knowledge Discovery from Data*, vol. 1, no. 5, 2007.

[9] P. Singla and P. Domingos, "Entity resolution with markov logic," in *Proceedings of the Sixth International Conference on Data Mining*, ser. ICDM. IEEE Computer Society, 2006, pp. 572–582.

[10] M. Wick, S. Singh, and A. McCallum, "A discriminative hierarchical model for fast coreference at large scale," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ser. ACL, vol. 1, 2012, pp. 379–388.

[11] M. Wick, A. Kobren, and A. McCallum, "Large-scale author coreference via hierarchical entity representation," in *Proceedings of the 30th International Conference on Machine Learning*, ser. JMLR: W&CP, vol. 28. Proceedings of the 30th International Conference on Machine Learning, 2013.

[12] P. Erdős and A. Rényi, "On the evolution of random graphs," *Magyar Tud. Akad. Mat. Kutató Int. Közl*, vol. 5, pp. 17–61, 1960.

[13] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.

[14] B. Bollobás, *Random Graphs*. New York: Cambridge University Press, 2001.

[15] D. Blackwell and J. B. MacQueen, "Ferguson distributions via Polya urn schemes," *Ann. Statist.*, vol. 1, no. 2, pp. 353–355, 1973.

[16] G. U. Yule, "A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f.r.s," *Phil. Trans. Royal Soc. B*, vol. 213, pp. 21–87, 1925.

[17] S. Goldwater, T. L. Griffiths, and M. Johnson, "Interpolating between types and tokens by estimating power-law generators," in *Advances in Neural Information Processing Systems 18*, 2005, pp. 459–466.

[18] J. Pitman and M. Yor, "The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator," *Ann. Prob.*, vol. 25, no. 2, pp. 855–900, 1997.

[19] J. Ferry, "Exact Association Probability for Data with Bias and Features," *Journal of Advances in Information Fusion*, vol. 5, no. 1, pp. 41–67, 2010.

[20] R. Popp, K. Pattipati, and Y. Bar-Shalom, "m-best S-D Assignment Algorithm with Application to Multitarget Tracking," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 37, no. 1, pp. 22–39, Jan. 2001.

[21] R. Jonker and A. Volgenant, "A Shortest Augmenting Path Algorithm for Dense and Sparse Linear Assignment Problems," *Computing*, vol. 38, pp. 325–34, 1987.

[22] M. E. J. Newman, "Random graphs as models of networks," in *Handbook of Graphs and Networks*, S. Bornholdt and H. G. Schuster, Eds. Weinheim, Germany: Wiley-VCH, 2003, pp. 35–68.

[23] A.-L. Barabási, *Linked: How Everything Is Connected to Everything Else and What It Means*. Plume, 2002.

[24] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, pp. 98–101, 2008.

[25] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75–174, 2010.

[26] M. B. Hastings, "Community detection as an inference problem," *Phys. Rev. E*, vol. 74, no. 3, p. 035102, 2006.

[27] J. P. Ferry and J. O. Bumgarner, "Tracking group co-membership on networks," in *Proc. 13th Int. Conf. on Information Fusion*, July 2010.

[28] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, "Inference and phase transitions in the detection of modules in sparse networks," *Phys. Rev. Lett.*, vol. 107, p. 065701, 2011.