ENRES: A Semantic Framework for Entity Resolution Modelling

Bradley Malin and Latanya Sweeney

November 2005
CMU-ISRI-05-134

Data Privacy Laboratory
Institute for Software Research International
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Entity resolution, the process of determining if two or more references correspond to the same entity, is an emerging area of study in computer science. While entity resolution models leverage artificial intelligence, machine learning, and data mining techniques, relationships between various models remain ill-specified. Despite growth in both research and literature, investigations are scattered across communities with minimal communication. This paper introduces a conceptual framework, called ENRES, for explicit and formal entity resolution model definition. Through ENRES, we illustrate how several models solve related, though distinctly different, variants of entity resolution. In addition, we prove the existence of entity resolution challenges yet to be addressed by past or current research.

Keywords: entity resolution, link analysis, concept graphs, record linkage, deduplication, object identification

1 Introduction

For over a decade, computer science has been evolving from a set of theoretical and basic engineering challenges towards the incorporation of intrinsically complex social, organizational, and political environments in which computers are situated. An indication of this shift is observable in the swell of research literature covering investigations into complex social and relational systems, such as collaborative filtering, link analysis, and social networks. While techniques and theories from prior computer science research in more traditional topics with mature foundations, such as graphical models and computer networks, can be effective for modelling and accounting for social systems, but such theory can not be blindly relied upon. The similarities may be no more than superficial; it is unclear if accepted theories from standard areas translate into the social setting. It is clear that the construction of a formal computational basis for modelling and incorporating social theory into traditional structures is a necessary direction in computer science research.

Of particular interest is the degree to which social and relational information can be data mined for interesting patterns. The relationships which are learnable from complex relational data cover a vast range of concepts, such as the discovery of clusters of similar entities in their personal likes/dislikes, the prediction of purchasing habits, and the collection and modelling of social structures for intelligence agencies. One of the more fundamental relationships is the linkage, or merging, of information corresponding to the same entity. The ability to determine when multiple pieces of data correspond to the same entity is crucial to a wide range of critical data mining and management processes, including data fusion, cleaning, and profiling. A number of computer science communities have investigated this notion under various names, such as "record linkage" [40], "deduplicaton" [37], "object identification" [30], and "word sensing" [29]; but each tends to design methods tailored to their own perceived challenges. Yet, as the number of communities studying this topic grows, in the literature there is an emerging notion of a common concept, recently dubbed "entity resolution" [2, 12].

The entity resolution problem can be informally defined as follows. Imagine there exists a set of entities, such as locations, people, or definitions. A recipient is provided with a set of references to the entities, but not the mapping of reference to entity. The goal of entity resolution is to correctly reconstruct this mapping. To facilitate this process, research and systems developed for data warehousing and relational database management have produced sound architectures for storage, relational modelling, retrieval, and the aggregation of mass amounts of entity-specific data. Yet, traditional data management models tend to concentrate on databases where schemas are fully specified, or fuzzy relational schemas are supplied by a user or learned from the databases attributes [2, 13, 24, 25]. Given the complexity and distribution of the environments in which data now resides, it is difficult to apply or adapt traditional data integration techniques for entity resolution applications. More specifically, current methods for database schema matching are time consuming, error prone, and subject to semantic constraints which need to be supplied on a case-by-base basis, and as such do not scale well in large distributed environments.

Furthermore, the communities in which entity resolution is addressed cover a vast spectrum of ideologies and methodologies. As a result, the success of an entity resolution method is often dependent on assumptions well-known in one community, but not clearly understood or specified in another. The same method can be applied to different communities' problems in a mathematical or algorithmic sense, however, assumptions incorporated into the design process can limit a method's capability. Thus, when methods designed in one community are applied to problems studied in other communities, they can provide subpar results in comparison to methods designed by the importing community. Subsequently, this can lead to conflicting claims of method superiority which are difficult to validate and generalize to a broader context beyond the confines of a specific community.

A principle confounder is the failure to model the type of data utilized in the resolution process. Automated methods and algorithms for matching entity-specific data have existed since the 1950's [32], but the proliferation of low cost collection and storage technologies have facilitated the construction of datasets corresponding to a wide range of semantic features. For instance, the original methods for record linkage were based on string comparison of names [20] for tracking an individual's records over multiple collections, but personal information relating to one's self (*Who is the entity?*) is only one type of knowledge. Yet, other types of information is now utilized for entity resolution such as location-based information (*Where is the entity?*), or social-based information (*Who does the entity know?*), each of which carry different semantics and can influence the way resolution is achieved.

This paper introduces a simple framework, ENRES, for specifying entity resolution models. The framework makes explicit the assumptions and semantics utilized by various communities. As a logical system with defined parameters, it facilitates formal reasoning and proofs regarding components of the entity resolution problem. We demonstrate how ENRES can represent prior and current research models, as well as how such models relate and differ. Furthermore, ENRES proves the existence of both current entity resolution models, as well as a substantial number of which are open problems yet to be studied by any community.

2 ENRES Framework

The ENRES framework makes assumptions and necessary conditions regarding entities, and references to entities, explicit. As a result, the framework can be used a formal reasoning tool.

2.1 Framework Basics

The basic concepts are drawn from set theory. First, in Definition 1, we introduce entities and entity sets which are underlying phenomena.

Definition 1 (Entity / Entity Set) An entity is a unique and discrete element of a population P. An entity set E is a set of entities drawn from P, such that $\forall x, y \in E, x \neq y$. □

Entities are not necessarily observable, but we do observe references to entities. In ENRES, references, as specified in Definition 2, are observed as tuples over a set of attributes. Each attribute is a semantic category of information.

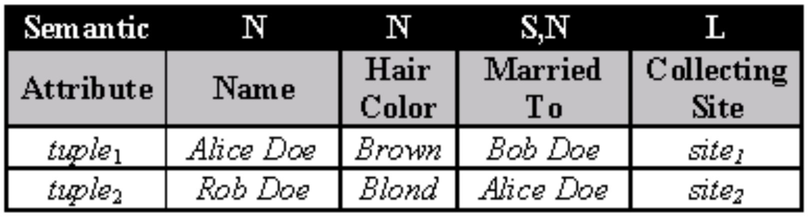
Definition 2 (Tuples / Tuple Set) Let A = $\{A_1, \ldots , A_n\}$ be a set of attributes. An n-tuple (or tuple) $t[a_1, \ldots , a_n]$ is a reference to an entity, such that $a_1 \in A_1, \ldots , a_n \in A_n$. A tuple set T_A is a set of tuples defined over A. □

In the current ENRES framework, attributes are partitioned into three types of semantics, dependent on the information they communicate. In general, dependencies take the form of 1) personal (N), 2) locational (L), and 3) social (S) and are more specifically defined in Definition 3. Several examples follow from Figure 1.

Definition 3 (Semantic Attribute) An attribute A is of semantic dependency type:

- *personal*, if A refers to the entity itself,
- *locational*, if A refers to locations where data is collected, and
- *social*, if A refers to relationships between entities. □

In Figure 1, the attribute "Hair Color", is personal dependent. It does not specify where the entity was and who the entity knows. In contrast, "Collecting Site" is location dependent since it denotes where data was gathered. An attribute can simultaneously satisfy more than one semantic. This is exemplified by "Married To", in which there exists a social relationship between "*Alice Doe*" and "*Bob Doe*", while the name "*Bob Doe*" itself is personal dependent.



Semantic	N	N	S,N	L
Attribute	Name	Hair Color	Married To	Collecting Site
tuple$_1$	*Alice Doe*	*Brown*	*Bob Doe*	*site*$_1$
tuple$_2$	*Rob Doe*	*Blond*	*Alice Doe*	*site*$_2$

Figure 1: Sample attributes and tuples. Semantic attribute types are depicted in the top row.

Moreover, the ENRES framework permits specification of three types of relations: 1) tuple-only, 2) entity-only, and 3) tuple-entity. Definition 4 provides formal definitions of these relations. A relation which maps tuples to tuples is called a tuple-only relation, entities to entities an entity-only relation, and tuples to entities is a tuple-entity relation.

Definition 4 (Relation Type) Let E be an entity set. Let $\mathsf{T} = \{T_A, T_B, \ldots, T_Z\}$ be a set of tuple sets referencing E. Let $T^* = \bigcup_{T \in \mathsf{T}} T$. A relation r is:

- $entity-only$, if $\forall(x, y) \in R$, $x \in E$ and $y \in E$,
- $tuple-only$, if $\forall(x, y) \in R$, $x \in T^*$ and $y \in T^*$,
- $tuple-entity$, if $\forall(x, y) \in R$, $x \in T^*$ and $y \in E$. □

Certain relations may be dependent on attribute semantics. In this sense, specifications on the relation are tantamount to necessary conditions. When dependency exists, the relation is represented with the superscript of the semantic types. For example, relation r^N acts on personal dependent attributes.

One particular tuple-entity relation represents the ground truth regarding tuples and the entities to which they correspond. This relation is called the *truth resolution function* and consists of the properties as laid out in Definition 5.

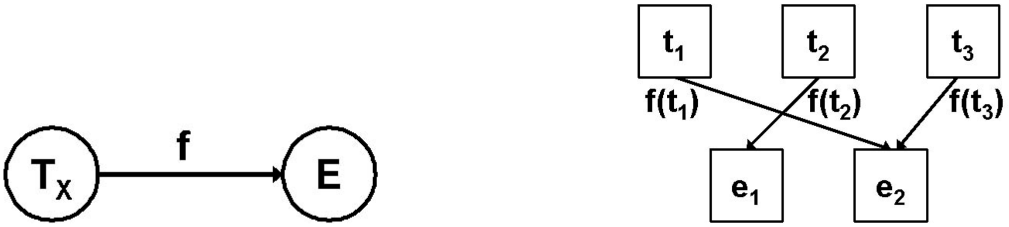
Definition 5 (Truth Resolution Function) Let E and T_A be a set of entities, such that T_A is representative of E. A function $f_A : T_A \rightarrow E$ is said to be truth resolution function, if it satisfies the following properties:

1. $\forall t \in T_A$: $\exists e \in E$, $f_A(t) = e$, and
2. $\forall e \in E$: $|{f_A}^{-1}(e)| > 0$. □

In combination, the first and second properties guarantee the truth resolution function is onto and many-to-one from tuples to entities. Yet, truth resolution functions may be unknown. Therefore, we introduce the concept of an *approximate resolution relation*. This relation maps tuples to entities, but permits non-unique resolution. The approximate resolution relation is dependent on known tuple-only and entity-only relations.

2.2 ENRES in Concept Graphs

ENRES can be specified in a graphical form. In the graph setting, as depicted in the left of Figure **??**, circular nodes correspond to sets of entities or tuples and edges correspond to relations between nodes. More detailed, we model individual elements of a set, in which square nodes are used, as shown in the right of Figure **??**.



(a) General Map

(b) Specific Map

Figure 2: Basic truth function mapping. 2(b) For tuple set to entity set. 2(b) For specific tuples to specific entities.

For approximate resolution, we use tuple-only, entity-only, and tuple-entity relations. An example of approximate resolution is depicted in Figure 3, which represents a resolution model for the truthful graph in Figure **??**. The truth resolution function f is searched for via the approximate resolution relation g, which itself is dependent on the tuple-only relation $h^{N,S}$. In Figure **??**, the latter relates one tuple set to itself via entity and social dependent attributes.

3 Entity Resolution Variants

In the following sections, we illustrate the viability of ENRES by surveying and modelling certain sections of the entity resolution research landscape.

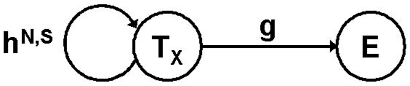


Figure 3: Approximate resolution example for with personal and social dependency.

3.1 Record Linkage and Deduplication

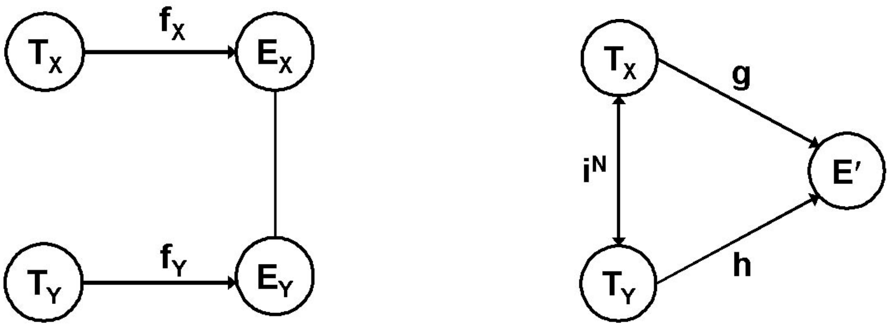
Some of the earliest research on the automation of entity resolution dates back to the middle of the twentieth century. At this time, Newcombe et al. [32] introduced methods for linking files from one database to another for ”record linkage”. The goal was to merge two lists **X** and **Y** of vital (i.e. health) information based on personal and demographic attributes, such as name, date of birth, gender, and residential address. One of the main assumptions of their model was files have common variables and there exists typographical error (e.g. “John Smith” vs. “Jon Smith”) or alternate representation of entities (e.g. “Rob” vs. “Bob”). They proposed an automated method for linking records which were similar in values. Names were stemmed and normalized using the Soundex coding scheme and a simple heuristic was used to score the probability that two records should be matched or not.

A statistical basis for record linkage was developed by Fellegi and Sunter [11] who introduced more formal decision criteria. Their interest in the problem was motivated by how to link census and governmental databases. Their method consisted of building a statistical model to classify pairs from the product space $\mathbf{X} \times \mathbf{Y} \rightarrow \{M, U, C\}$, where M is the set of definite matches, U is the set of definite non-matches, and C is a set of pairs that need clerical review. The Fellegi-Sunter methods were ushered into the modern statistical age by Winkler [40, 41] who demonstrated how iterative expectation-maximization methods could be used to improve upon the original static methods. More recently, Pasula et. al. [34] investigated record linkage as the “identity uncertainty” problem. In this latter work, probabilistic relational models are adapted for resolving uncertainties in the author names and titles of paper citations.

To assist modern models, Soundex-based methods have been improved upon for string comparison and distance metrics for record linkage based on personal dependent strings [5, 7]. Furthermore, researchers in the medical informatics community have evaluated the degree to which entity dependent attributes collected at hospitals, such as and names, dates, and Social Security Numbers, are stable and unique for linkage purposes [16, 39]. For instance, research by Sweeney demonstrated that demographic features, such as the combinations of values from Birthdate, Gender, 5-Digit Zip Code could be used to link medical and voter registration records for approximately 87% of the United States Population. [39]

In order to map record linkage to the ENRES framework, we represent **X** and **Y** as tuple sets T_X and T_Y. ENRES models truth resolution functions f_X and f_Y which map the tuple sets to their respective entity sets as E_X and E_Y. Note, in the underlying system the mapping between entities in E_X and E_Y is known. Figure 4 depicts this system.

Let x and y be tuples in T_X and T_Y, respectively. In terms of entity resolution, a correct match can be represented as $e = f_X(x) = f_Y(y)$, where $e \in E$. The specific problem that record linkage addresses is as follows. Imagine that f_X, f_Y, and E exist, but are unknown. It is known there exists some set of relations between T_X and T_Y over the set of personal dependent attributes N, which we will refer to as i^N. The goal is to discover approximate resolution relations g_X, h_Y and an entity set E', such that $e' = g(x) = h(y)$ only if $e = f_X(x) = f_Y(y)$.



(a) Truth resolution function

(b) Approximate resolution

Figure 4: Record Linkage. 4(b) Truth resolution function. 4(b) Approximate resolution relations.

In addition to linking two separate files, the underlying ideas of record linkage have also been referred to as record deduplication. The term deduplication corresponds to linking a single database to itself to remove records on the same individual which appear different. Some researchers contend that deduplication is equivalent to record linkage, where $T_X = T_Y$ and $g = h$. However, the ENRES framework suggests otherwise. In ENRES, the record linkage and deduplication problems are modelled in Figures 4 and 5, respectively. It is interesting to note that while the same statistical procedures can be applied to record linkage and deduplication, there is a fundamental difference in the problems. Specifically, when performing deduplication it is known that the set of entities which are referenced from the set of tuples is equivalent. Thus, there clearly exists two onto mapping functions. In contrast, when performing record linkage, the set of entities of two different tuple sets are not necessarily onto functions. This is because neither set of tuples is guaranteed to be onto the set of entities E.

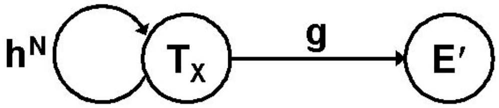


Figure 5: Record deduplication and approximate resolution relation.

3.2 Location Based Linkage

While record linkage and deduplication models are designed to account for personal dependent attributes over disparate databases, location-based linkage [28, 35, 38] resolves tuple-entity relations when such relations are missing. In contrast, for resolution, location-dependent attributes can be exploited to discover patterns in the locations where an entity's data was collected or originated.

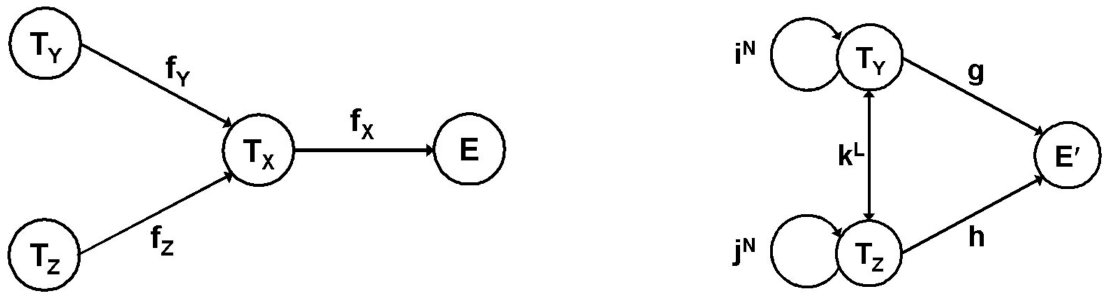
An example of location dependent linkage is the trail linkage model [28], which is based upon the observation that people visit different sets of locations where they leave behind data of type X.

Each visited location collects and, subsequently, shares data type X as two types of data Y and Z that can not be related via their personal dependent attributes. However, data of type Y and Z are traceable can be self related via personal dependencies (i.e. Y consists of personal demographics). When multiple locations share data, this allows for trails or characterizations of the locations that entities visited to be constructed. As a result, similar location visit patterns in the trails of type X and Y can be used for linkage purposes.

The problem of trail linkage can be expressed in terms of the ENRES framework as follows. Let E be a set of entities. Let T_X be a set of tuples representative of E, such that $X = \{X_1, ..., X_n\}$. Let f_X be a truth resolution function from T_X to E. Let Y and Z be sets of attributes such that:

1. $Y \subset X, Z \subset X$,
2. $|Y \cap Z| > 0$, and
3. ($Y \cap Z$) is of type L.

Now, consider two new sets of tuples T_Y and T_Z. Let f_Y be a function in which $\forall y \in T_Y, \exists x \in T_X, f_Y(y) = x$. Similarly, let f_Z be a function in which $\forall z \in T_Z, \exists x \in T_X, f_Z(z) = x$. Let $e = f_X(f_Y(y)) = f_X(f_Z(z))$. If E, T_X, f_X , f_Y, and f_Z are unknown, find relations g and h, such that $|g^{-1}(y)|$ = $|{f_Y}^{-1}(y)|$ and $|h^{-1}(z)|$ = $|{f_X}^{-1}(z)|$. By discovering such relations, we correctly link all pieces of information corresponding to the same entity.



(a) Truth resolution function

(b) Approximate resolution

Figure 6: Location-based Linkage. 6(a) The underlying truth resolution function. 6(b) The observed relationships for approximate resolution via trail linkage.

The discovery of relations g and h are aided, and to a certain extend defined, by imposing constraints or assumptions on the attributes which relate T_Y and T_Z. For trail linkage, The necessary conditions, or constraints, are required on the relationships between the observed sets of tuples:

1. $\exists$ tuple-only relation $i^N \subseteq T_Y \times T_Y$,
2. $\exists$ tuple-only relation $j^N \subseteq T_Z \times T_Z$, and
3. $\exists$ tuple-only relation $k^L \subseteq T_Y \times T_Z$.

Several variants of the trail re-identification problem which specify the relationships permitted between T_X, T_Y, and T_Z have been introduced [26]. In prior research, we posed several deterministic solutions tailored to specified assumptions collectively termed the REIDIT (RE-identification of Data In Trails) algorithms [26, 28]. Where statistical record linkage methods attempt to maximize the probability of a linkage, the REIDIT algorithms guarantee correct linkages when certain assumptions over the data hold true. One of the drawbacks to the methodology employed by REIDIT is that it simplifies the relations i^N and j^N to equivalence relations where data of the same type are considered to belong to have been generated by the same entity if they are equivalent in their entity-dependent values. Regardless, some the modelling and investigative techniques employed by the REIDIT algorithms may assist in the proposed work below. Yet, this model provides a foundation for exploiting simple relationships and understanding the basis regarding how the distribution of data affects for resolution goals.

3.3 Social Linkage and Deduplication

More recently, interaction and associations between entities in the form of social networks have been explored to solve variations of the entity resolution problem. Social interaction is observable when entities from the entity set are involved in some type organizational relationship, such as researcher co-authorships or communication networks of terrorists.

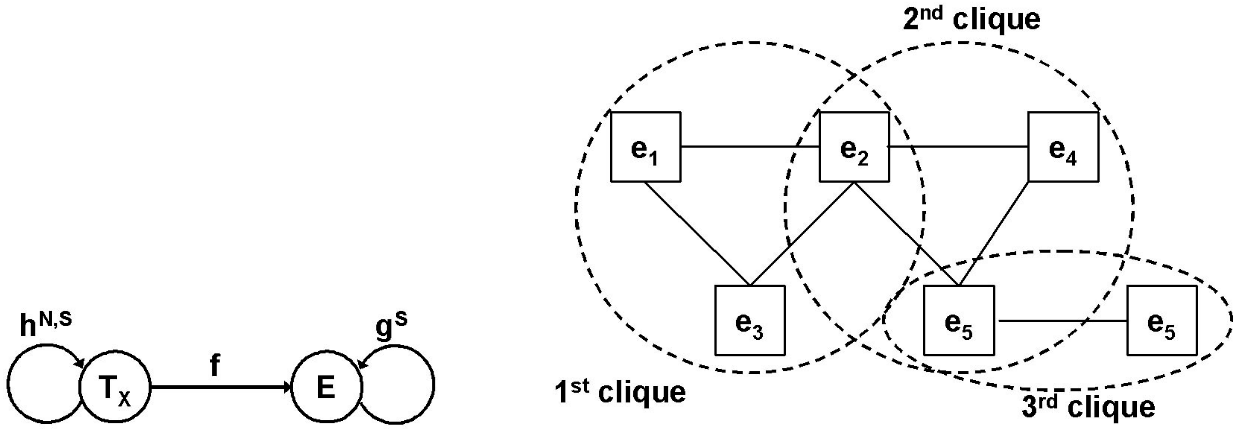
3.3.1 Deduplication in Labelled Networks

A network is considered labelled if its nodes provide personally identifying attributes. With respect to labelled networks, one area of entity resolution research has studied the interactions among co-citations and author collaborations. Imagine there exists a set of papers, the authors of which are drawn from a set of entities. How can we determine which papers were written by the same authors? Again, an author's name can vary (e.g. "Robert" versus "Rob") and the same name may correspond to multiple authors. We could study the writing styles used in the research papers, or possibly the topics of interest expressed, but the goal of social based entity resolution is to use the groups of names occurring together in such documents. [3, 4, 21, 19, 27]

To consider one specific case, in the models in [3, 4], assumptions are imposed on the entity-to-entity relationships. The main assumption is that subsets of an entity set E exist, though are unknown, in the form of cliques. A clique is defined as a set of entities $Q \subseteq E$, such that every pair of entities $e, f \in Q$ have a positive, or non-null, social relationship.

These relationships are observed in the set of names which appear as co-authors for a particular paper, the set of such papers make up T_X. These cliques do not manifest in perfect representation in the author list of a paper, since there is variation in the name of entities, and sometimes entities from outside of a clique are included in the local network for a collaboration. However, the authors impose an assumption that cliques are recoverable by partitioning papers, or sets of names, into groups which maximize clique-like phenomena. Paper grouping allows for the prediction of when authors names on disparate papers correspond to the same underlying entity.

In ENRES, the necessary conditions for this model is exemplified in Figure 7. Note, in this representation, the necessary condition for the relationship between T_X to T_X, is a social relationship. In other words, only groups of tuples are necessary. In [4, 3], string matching algorithms,



(a) Approximate resolution

(b) Imposed Model

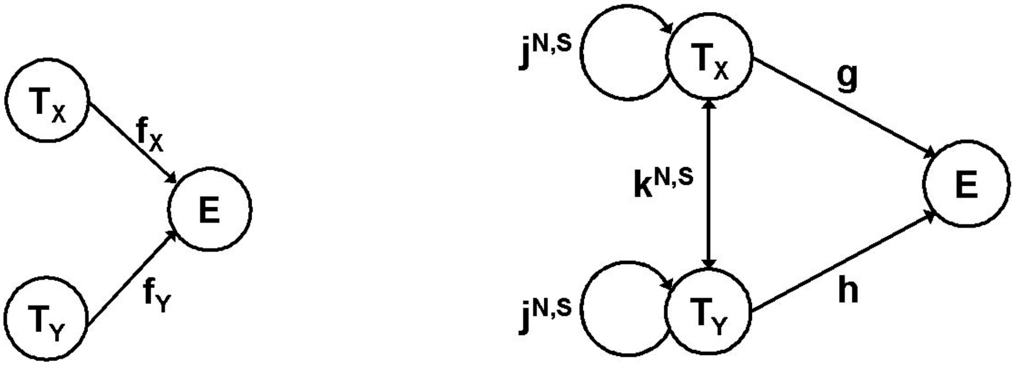
Figure 7: Labelled social deduplication. **??** Approximate resolution for labelled social deduplication. 7(b) Assumed social relationship model among entities.

which process personal-dependent attributes, are suggested for the resolution of noise in the name representations of the underlying entities. In this sense, the social network analysis functions as an additional feature by which deduplication occurs.

The specification of clique detection allows for a robust statistical learning model to be imposed on labelled datasets. Yet, there exists a tradeoff in specification of the underlying entity-entity relationships and the generalizability of learning. For instance, clique detection requires what we informally term exact similarity, such that relationships between entities must be directly observed (e.g. Alice and Bob are related if they collocate in the same source). This model is not necessarily representative of the space of social networks and it is unclear if this model generalizes to other types of social networks [1, 33], such as smallworld [22], hierarchical [36], or cellular [8]. Recently, it has been shown that such assumptions biases the learning realm and can have serious difficulty in lesser connected, or decentralized, environments such as the actor-to-actor relationships in the Internet Movie Database. [27]. Alternative learning models, which relax the underlying relationships to lesser structured systems, such as those based on network walks [21, 27], covering [10], cuts [14, 31], and spectral clustering [19] may provide intuition into additional social environments.

3.3.2 Linkage in Partially Labelled Networks

A related, though distinct, problem regarding social networks for resolution is the topic of link completion. [15, 23] This problem can be defined as follows. A social network is observed, where edges in the network denote the affinity to which two different nodes are related. The network can be constructed from a set of tuples T_X as described above. Then, the observer is presented with a new network, where the label of one node x is obscured. The set of such "partial" networks can be thought of as T_Y. The population of entities is closed, so the truth resolution functions for T_X and T_Y map back to the same entity set. The goal is to link the unlabelled node to its corresponding labelled node in the other network. In ENRES terms, this is equivalent to linking x to its corresponding entity, which itself is labelled by one or more tuples from T_Y. It is expected that both T_X and T_Y correspond to the same entity set as shown in the right of Figure 8. It is for this reason that in the resolution model depicted in the left of Figure 8 the entity set is E and not E' as was the case in the record linkage models described earlier.



(a) Truth resolution

(b) Approximate Resolution

Figure 8: Partially labelled social linkage model. 8(a) Underlying truth resolution functions. 8(b) Approximate resolution as linkage in partially labelled networks using social dependent relations.

An alternative view, though equivalent representation within the framework, is offered by Hill and Provost. [17, 18] In this scenario, the social network is a citation network where the authors of a paper are hidden from view. Instead of studying the groups of co-authors, the social network is constructed from the paper's citation list. Resolution on a paper with an unknown author is achieved via a classifier trained on observed citation networks for known author.

By comparing Figures 7 and 8, it is apparent that the clique detection research is represented in the framework's form of deduplication with different constraints on the resolving relationships. In contrast, in the classification setting, the problem appears to be related to the linkage problems of trail and record linkage. More specifically, as depicted in 8, the goal is to link a newly observed network to one of a set of networks observed in the construction of the classifier.

4 Discussion

The presented models merely scratch the surface of the entity resolution landscape. One strength of ENRES resides in its ability to formally model the structure of entity resolution problems. In this section, we begin to characterize how many resolution models can exist. Then, we propose a new entity resolution problem yet to be addressed by any research community was discovered via ENRES modelling.

4.1 Many Resolution Models

The graphical nature of ENRES supports an algebraic investigation of model topologies. We assume the entity set population is known.

4.1.1 Deduplication

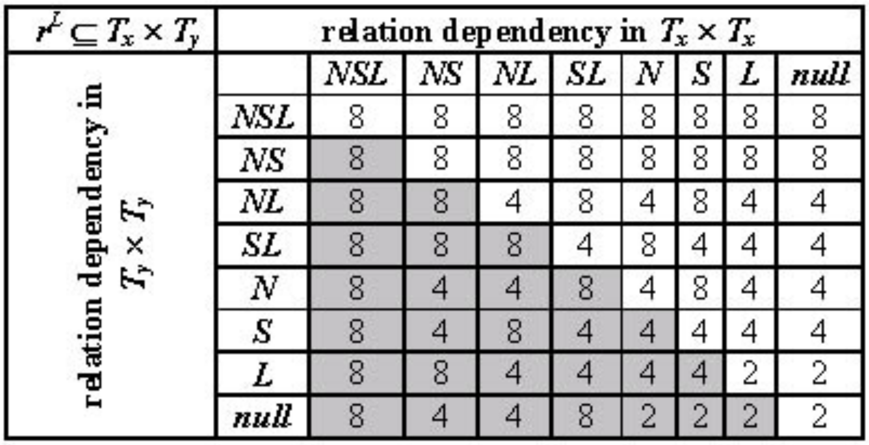
For deduplication there is one tuple set and one entity set. A tuple-only relation exists and is dependent on at least on semantic. The number of such relations is 2^{n-1}, where n is the number of semantic types. In addition, entity-only relations are not required, so 2^n, or 8 are possible. However, it is counterintuitive to make entity-only relations dependent upon semantics which are unobservable. As a result, entity-only semantics are selected from the set used for tuple-only relations. The number of models can be computed as

$$\sum_{i=1}^{n} \binom{n}{i} 2^i.$$

Consequentially, when $n = 3$, the number of deduplication models is 26.

4.1.2 Linkage

We consider the case of two tuple sets T_X, T_Y and one entity set E. The only requirement is the relation between T_X and T_Y (i.e. $r \subseteq T_X \times T_Y$) has at least one semantic dependency. There are $2^n - 1$ such relations. In addition, each tuple set can be self-related using 2^n possible semantic combinations. This is depicted along the top row and first column of Figure 9. Next, the number of entity-only relations is dependent on the semantics of the tuple-only relations. In Figure 9, when $r \subseteq T_X \times T_Y$ is locational dependent, this number is the sum of the diagonal plus the right upper triangle. This matrix sums to the same value when $r \subseteq T_X \times T_Y$ is dependent on one semantic. Similar matrices can be constructed for any number of semantic dependencies for $r \subseteq T_X \times T_Y$.



$r^L \subseteq T_x \times T_y$	relation dependency in $T_x \times T_x$							
relation dependency in $T_y \times T_y$		***NSL***	***NS***	***NL***	***SL***	***N***	***S***	***L***	***null***
	NSL	8	8	8	8	8	8	8	8
	NS	8	8	8	8	8	8	8	8
	NL	8	8	4	8	4	8	4	4
	SL	8	8	8	4	8	4	4	4
	N	8	4	4	8	4	8	4	4
	S	8	4	8	4	4	4	4	4
	L	8	8	4	4	4	4	2	2
	null	8	4	4	8	2	2	2	2

Figure 9: Number of different entity-only relations given when $T_X \times T_Y$ is locational dependent. The matrix is symmetric so gray cells are redundant.

When $n = 3$ and the number of semantics for $r \subseteq T_X \times T_Y$ is 2, there are 36 possible combinations of self referential tuple-only relations, 27 of these do not contribute the missing semantic. Furthermore, when the cross tuple relation contributes one semantic, 19 of the 36 combinations communicate both missing semantics and 14 of the 36 communicate one of the missing semantics. Thus, when n=3, the number of possible models can be calculated as 36*8 + 3*(27*8 + 9*4) + 3*(19*8 + 14*4 + 3*2) = 1686.

4.2 A New Model: Topological Linkage

In the previous entity resolution investigations with social networks, the question of interest was "Given who x interacts with, can we determine who x is?" In this setting, we resolve who someone is given their labelled interactions. Yet, a problem not addressed by such research is how to handle situations when all nodes in the network are unlabelled. While it may be known that interactions exist, it is not known who (or some entity-dependent reference of who) those entities are. In this setting, the question "Given what someone's social network looks like, but not the identities of the network, can we determine who that someone is?" is more appropriate. The goal is to link nodes from a social network with no explicit identities to corresponding nodes in labelled social networks. We term this variant of the entity resolution problem *topological linkage*.

The topological linkage problem occurs in many real world situations, including privacy and re-identification analysis, fraud detection, and covert network analysis. The latter is quite interesting to note. In prior covert network analysis research, the goal is not the resolution of individual nodes, but instead resolution of the entire network. For example, it is not so important to know exactly who a particular node is, but what faction or group the network, which the node resides in, represents. In this respect, covert network analysis methods attempt to discover the organizational structure of the network in T_Z. [6, 8, 9]. This too is an entity resolution problem, but at a macro scale, whereas in the problem of topological linkage the goal is to link each specific unlabelled node to a specific labelled node.

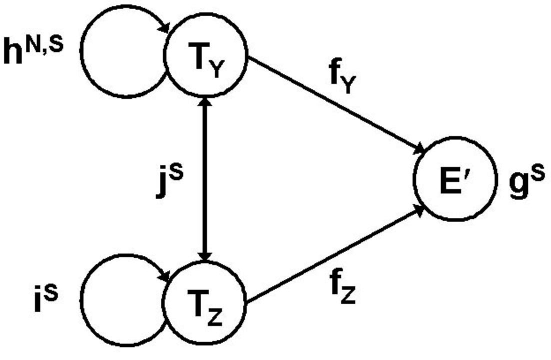


Figure 10: ENRES topological linkage model.

The ENRES model of topological linkage is shown in Figure 10. The underlying truth is the same as that depicted in Figure 8. In this problem, we are provided with, or construct, a social network from the set of tuples T_Z. This is an unlabelled network and none of the nodes are required to be labelled with personal-dependent attributes. The goal is to link, or label, the nodes through observable interactions, and the subsequently constructed network, of entities in a labelled setting. The latter labelled network is derived from the set of tuples T_Y and is organized through relation $h^{N,S}$.

4.3 Limitations and Extensions

The ENRES model is limited in certain respects, several of which we address here. First, the semantic types ENRES utilizes are derived from surveys of research literature related to entity resolution. As such, the specification of three types is arbitrary. Yet, the separation of data types serves to provide a first approximation of model semantics. It is possible that additional semantics exist and can be integrated to make ENRES more robust. One direction for our future research is to derive data semantic types which are less dependent on survey and more dependent on formal characteristics.

A second limitation of ENRES derives from its lack of decision support. A researcher can model data semantics and the resolution problem, but ENRES does not explicitly provide feedback to the researcher regarding which methods are best suited to solve the resolution problem. We believe this is a logical extension to the ENRES framework. Since ENRES uses a logical structure for resolution models, it is not difficult to converted into a case-based reasoning system. For instance, ENRES could be trained with samples of ¡model, method¿ pairs, such that when a new model is presented, ENRES predicts which method(s) is best suited, or most probable, to achieve resolution.

5 Conclusions

This paper introduced a framework for entity resolution, ENRES, which provides a common architecture for modelling seemingly disparate research on how to determine if two pieces of data correspond to the same entity. Previous research into topics such as record linkage and link completion were mapped into the framework. Furthermore, we demonstrated that assumptions, such as semantic types of attributes, can be made explicit. In addition, we derived, via ENRES modelling, a new entity resolution problem called topological linkage, which is defined as linking specific nodes from unlabelled to a labelled social networks. ENRES sets the basis for a case-based reasoning tool for determining which methods are best suited to solve a given resolution problem.

6 Acknowledgments

The authors thank Kathleen Carley, Christos Faloutsos, and the members of the Data Privacy Laboratory for useful discussions, support, and precious time. This research was partially supported by the Data Privacy Lab at Carnegie Mellon University and NSF IGERT grant 9972762. The opinions expressed in this research are those of the authors and do not necessarily reflect those of the National Science Foundation.

References

[1] R. Albert and A. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[2] O. Benjelloun, H. Garcia-Molina, Q. Su, and J. Widom. Swoosh: a generic approach to entity resolution. Technical Report 2005-5, Stanford University, Palo Alto, CA, 2005.

[3] I. Bhattacharya and L. Getoor. Deduplication and group detection using links. In *Proceedings of the ACM Workshop on Link Analysis and Group Detection (LinkKDD-2004)*, 2004.

[4] I. Bhattacharya and L. Getoor. Iterative record linkage for cleaning and integration. In *Proceedings of the ACM SIGMOD on Research Issues in Data Mining and Knowledge Discovery*, pages 11–18, 2004.

[5] M. Bilenko and R. Mooney. Learning to combine trained distance metrics for duplicate detection in databases. Technical Report AI-02-296, AI Laboratory, Univerisity of Texas, Austin, TX, Feb 2002.

[6] K. Carley, M. Dombroski, M. Tsvetovat, J. Reminga, and N. Kamneva. Destabilizing dynamic covert networks. In *Proceedings of the 8^{th} International Command and Control Research and Technology Symposium*, Washington, DC, 2000.

[7] W. Cohen, P. Ravikumar, and S. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, 2003.

[8] M. Dombroski, P. Fischbeck, and K. Carley. Estimating the shape of covert networks. In *Proceedings of the International Command and Control Research and Technology Symposium*, Washington, DC, 2003.

[9] P. Drineas, M. Krishnamoorthy, M. Sofka, and B. Yener. Studying e-mail graphs for intelligence monitoring and analysis in the absence of semantic information. In *Proceedings of the Symposium on Intelligence and Security Informatics*, Tucson, AZ, 2004.

[10] C. Faloutsos, K. McCurley, and A. Tomkins. Fast discovery of connection subgraphs. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 118–127, 2004.

[11] I. Fellegi and A. Sunter. A theory for record linkage. *Journal of the Americal Statistical Association*, 64:1183–1210, 1969.

[12] H. Garcia-Molina. Entity resolution: Overview and challenges. In *Proceedings of the International Conference on Conceptual Modeling*, pages 1–2, Shanghai, China, 2004.

[13] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of link structure. *Journal of Machine Learning Research*, 3(4–5):679–708, 2003.

[14] M. Girvan and M. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, USA*, 99:7821–7826, 2002.

[15] A. Goldenberg, J. Kubica, P. Komarek, A. Moore, and J. Schneider. A comparison of statistical and machine learning algorithms on the task of link completion. In *Proceedings of the ACM SIGKDD Workshop on Link Analysis for Detecting Complex Behavior*, 2003.

[16] S. Grannis, J. Overhage, and C. McDonald. Analysis of identifier performance using a deterministic linkage algorithm. In *Proceedings of the American Medical Informatics Annual Symposium*, pages 305–309, 2002.

[17] S. Hill. Social network relational vectors for anonymous identity matching. In *Proceedings of the IJCAI Workshop on Learning Statistical Models from Relational Data*, Acapulco, Mexico, 2003.

[18] S. Hill and F. Provost. The myth of the double-blind review?: author identification using only citations. *ACM SIGKDD Explorations*, 5(2):179–184, 2003.

[19] P. Hsiung, A. Moore, D. Neill, and J. Schneider. Alias detection in link data sets. In *Proceedings of the International Conference on Intelligence Analysis*, McLean, VA, 2005.

[20] M. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84:414–420, 1989.

[21] D. Kalashnikov, S. Mehotra, and Z. Chen. Exploiting relationships for domain-independent data cleaning. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 262–273, Newport Beach, CA, 2005.

[22] J. Klienberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the 32^{nd} Annual ACM Symposium on Theory of Computing*, Portland, OR, 2000.

[23] J. Kubica, A. Moore, D. Cohn, and J. Schneider. Finding underlying connections: A fast graph-based method for link analysis and collaboration queries. In *Proceedings of the International Conference on Machine Learning*, pages 392–399, Washington, DC, 2003.

[24] A. Laurent. Querying fuzzy multidimensional databases: unary operators and their properties. *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, 11:31–45, 2003.

[25] C. Li and X. Wang. A data model for supporting on-line analytical processing. In *Proceedings of the ACM Conference on Information and Knowledge Management*, Rockville, MD, 1996.

[26] B. Malin. Betrayed by my shadow: learning data identity via trail matching. *Journal of Privacy Technology*, page 20050609001, 2005.

[27] B. Malin. Unsupervised name disambiguation via social network similarity. In *Proceedings of the SIAM Workshop on Link Analysis, Counterterrorism, and Security*, pages 93–102, Newport Beach, CA, 2005.

[28] B. Malin and L. Sweeney. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomededical Informatics*, 37(3):179–192, 2004.

[29] G. Mann and D. Yarowsky. Unsupervised personal name disambiguation. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 33–40, Edmonton, Canada, 2003.

[30] M. Neiling and S. Jurk. The object identification framework. In *Proceedings of the ACM SIGKDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, pages 33–40, Washington, DC, 2003.

[31] J. Neville, M. Adler, and D. Jensen. Clustering relational data using attribute and link information. In *Proceedings of the IJCAI Workshop on Text Mining and Link Analysis*, Acapulco, Mexico, 2003.

[32] H. Newcombe, J. Kennedy, S. Axford, and A. James. Automatic linkage of vital records. *Science*, 130:954–959, 1959.

[33] M. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[34] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. Identity uncertainty and citation matching. In *Proceedings of Neural Information Processing Systems*, Vancouver, Canada, 2003.

[35] N. Priyantha, A. Chakraborty, and H. Balakrishnan. The cricket location-support system. In *Proceedings of the ACM Inferenational Conference on Mobile Computing and Networking*, pages 32–43, Boston, MA, 2000.

[36] E. Ravasz and A. Barabasi. Hierarchical organization in complex networks. *Physical Review E.*, 67:026112, 2003.

[37] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Vancouver, Canada, 2002.

[38] A. Smailagic and D. Kogan. Location sensing and privacy in a context-aware computing environment. *IEEE Wireless Communications*, 9:10–17, 2002.

[39] L. Sweeney. Uniqueness of simple demographics in the u.s. population. Technical Report LIDAP-04, Data Privacy Laboratory, CMU, Pittsburgh, PA, 2000.

[40] W. Winkler. Matching and record linkage. In B. Cox, editor, *Business Survey Methods*. Wiley, New York, NY, 1995.

[41] W. Winkler and F. Scheuren. Recursive analysis of linked data files. In *Proceedings of the Census Bureau Annual Research Conference*, pages 920–935, Washington, DC, 1996.