WHITE PAPER

HOLISTIC IDENTITY RESOLUTION

John R. Talburt, PhD
Director, Center for Advanced Research in Entity Resolution and Information Quality (ERIQ), University of Arkansas at Little Rock

Richard Y. Wang, PhD
Director, Chief Data Officer & Info Quality Program at Massachusetts Institute of Technology (MIT)

Dirk Beyer, PhD
Senior Vice President and General Manager, OneID and Data Science, Neustar

TABLE OF CONTENTS

THE CHANGING LANDSCAPE OF IDENTITY RESOLUTION

Customer relationship management (CRM) has traditionally relied on the process of "customer recognition," the application of entity and identity resolution to customers. Customer recognition hinges on the premise that the first step in establishing a customer relationship is to resolve the identity of the customer through his or her channel of contact.

A customer is an example of an entity. In the context of entity and identity resolution, entities are real-world objects distinguishable from each other by a unique set of characteristics or attributes. These attributes often take the form of identifiers for the entity. A person usually has several identifiers to establish their identity in different contexts. A name, address, phone number, and email address can all be used to identify a person in the physical world, while a mobile ad ID (MAID), browser cookie ID, IP address, and device ID can be used to identify a person in the digital world. Each identifier has a different shelf life, and the combination of these signals or IDs are used to determine a person's identity. The accuracy, relevancy, and stability of that identity depend on exactly how these identifiers are linked.

The challenge of identity resolution in a business context is that we are often limited to merely inferring the unique identity of a customer, by collecting and storing a sufficient combination of these identifiers. The goal of an identity resolution system is to create a distinct representation of every customer and assign a unique identifier to each one. To be effective though, the customer identifier assigned by your identity resolution system needs to be persistent. While individual customer identifiers may change from time to time, such as a change of address or change of name, the system-assigned identifier for the customer should not change. The inability to consistently identify the same information from different sources of data is one the leading causes of data quality problems in organizations.[1]

Customer recognition systems were some of the first identity resolution systems developed. Most of these systems were based on relatively simple models that assumed in-store, direct mail, and telephone calls were the primary channels of interaction. At that time identity was much more transparent.

[1] *Journey to Data Quality*, Lee, Pipino, Funk, Wang, MIT Press

However, the challenge of identity resolution has dramatically increased since the days of direct mailing and telemarketing. Two of the most significant new challenges in identity resolution are: 1) the number of channels and touchpoints in which customers can be engaged has dramatically increased, and 2) there has been an industry shift in the way we define and use identity due to the digital and physical worlds becoming increasingly intertwined. Customers now assume many different personas spread across a multitude of devices, browsers, websites, and social media platforms. A 2014 study by the Aberdeen Group[2] found that, on average, a typical company interacts with a customer over nine different channels. For many organizations they found that the number exceeded 12 channels. Given the tremendous growth in social media, this average has no doubt increased since then. Facebook, Instagram, Pinterest, Snapchat, LinkedIn, Twitter, WhatsApp, WeChat—the list seems to grow longer every day. Not to mention the many popular online apps and on-demand services such as those for ride-sharing, lodging rentals, food delivery, and online shopping that have added additional pieces to the mosaic of a customer's identity. They provide not only contact signals, but also customer behavior, attitudes, preferences, and propensity.

Another change has been the rising importance of "personalization" across every customer experience. While the goal of identity resolution is to establish the true identity of the customer, personalization is often executed without knowing the identity of the customer. Many of us have experienced this firsthand when browsing for a particular item across the open web. For instance, shopping for a particular camera model on one website, and then later, while on a different website, seeing ads for similar or even the exact same camera model. This form of personalization or retargeting occurs when a marketer has collected product-level information via cookies from multiple website engagements, likely including the marketer or brand's website. This is all orchestrated without knowing the identity of the person, instead using the identity of the browser session in the form of a cookie.

The trend toward personalization has sharpened the contrast between **entity** resolution vs. **identity** resolution. Entity resolution simply determines when two references to real-world entities are referring to the same entity or different entities[3], a process of disambiguation. Identity resolution, however, determines whether an entity reference is referring to a known identity in a system of record, such as a customer profile within a CRM system, a process of recognition.

While the goal of an identity resolution system is to either determine a match of a collected entity reference to a stored identity or to create a new identity, some entities or signals may be too incomplete or ambiguous to do either with a high degree of confidence. Such incomplete and ambiguous references are called "identity fragments." Some identity resolution systems merely discard these fragments. However, if properly managed, these fragments can still provide value. In some cases, they can be used for simple personalization or retargeting as described above. Additional information may also be collected later, allowing the system to ultimately resolve the aggregate information to an identity cluster.

> The inability to consistently identify the same information from different sources of data is one of the leading causes of data quality problems in organizations.

[2] Minkara, Omer. *Customer Communication Management: Maximizing CEM Results with Interactive Content.* Aberdeen Group White Paper, Nov 2014.
[3] J. Talburt and Y. Zhou. *Entity Information Life Cycle for Big Data.* Morgan Kaufmann, 2015

HOLISTIC IDENTITY RESOLUTION

Traditional identity resolution systems employ one of two approaches to recognition, deterministic or probabilistic. Deterministic resolution uses conditional rules for matching decisions, whereas probabilistic resolution uses weights calculated from estimated probabilities to generate a numerical confidence score for its matching decisions. While both approaches have their advantages and disadvantages, by themselves they are not sufficient to cope with the rapid changes now occurring in customer identity, including identity fragmentation and the need for more complex forms of personalization.

The first reason they cannot cope is that most traditional identity resolution systems are designed to operate on standardized offline references. They are often "hard-wired" for only specific types of entities, such as person, household, and postal delivery point. These systems also usually store a fixed set of characteristics like name, address, and phone number to resolve the identity of these entities.

These limitations are often an artifact of the system's underlying implementation as a relational database management system (RDBMS). Relational databases are notorious for the rigidity of their data model and the difficulty of changing the model once it is in operation. Each entity is a table with a fixed number of characteristics (columns) with a foreign key used to provide a relationship link between two tables.

Most customer identity resolution systems still try to represent each customer entity as a single "golden record" with a static set of identity attributes. While this master data management (MDM) approach can be effective for highly-structured, offline data, it is not well suited for combining online and offline references, nor is it suitable for preserving and organizing identity fragments for later identification. More robust and flexible systems are required to capture and manage the vast and diverse set of digital signals produced by today's omnichannel customers. The modern consumer leaves behind a wide array of digital breadcrumbs such as cookies, MAIDs, and IP addresses, as well as intent-related signals such as impressions, views, clicks, and panel-based metrics across addressable media (digital display, mobile web, mobile in-app).

The complexities of the customer journey require us to develop and evaluate new methodologies for identity and entity resolution, approaches going beyond traditional, static methods. Dramatic challenges call for a more holistic approach to identity resolution. Holistic identity resolution systems combine the best of traditional deterministic and probabilistic approaches with newer technologies and tools such as graph databases and machine learning.

Holistic identity resolution systems go beyond simple matching. For example, some approaches develop probabilistic evidence of identity from behavioral data such as repeated co-occurrences of cookies on the same network or on networks with a certain degree of proximity. Holistic systems weigh evidence across all channels simultaneously rather than performing a separate analysis within each channel and joining results. In this way, a holistic system can establish more linkages with higher accuracy.

The idea of holistic identity resolution is analogous to the work of an archaeologist reconstructing pottery from an ancient burial. The shards of many different pottery items lie intermingled in the same location. Each shard must be examined and carefully fitted with other shards sharing the same pattern. Some shards can be identified and fit together immediately, while other shards must be set aside in the hope that the pottery item to which it belongs will become more evident as the items are reconstructed. Two key features of holistic identity resolution are: 1) flexibility in structure and operation, and 2) agility in responding to change.

Holistic identity resolution systems combine the best of traditional deterministic and probabilistic approaches with newer technologies and tools such as graph databases and machine learning.

FLEXIBLE IDENTITY REPRESENTATION IN GRAPH STRUCTURES

"Identity graphs" are a specialized form of "knowledge graphs". Knowledge graphs have been an essential type of data structure in the field of artificial intelligence (AI) for many years. The knowledge graph began as a system for organizing metadata for the Internet. Google pioneered them to power its search engine, associating meaning (semantics) to web pages to create a machine-readable representation of the concepts they describe. They could then use this ontology to automate reasoning about queries related to these web pages.

Figure 1 shows a simple identity graph for two persons, John and Mary. The edges of the graph represent relationships such as "John – is married to – Mary" and "Mary – uses – IP.23.12.1." The nodes of the graph are various types of entities such as persons, locations, and devices. These types of entity-to-entity relationships are the basis for the Resource Description Framework (RDF), which supports the Web Ontology Language (OWL)[4].

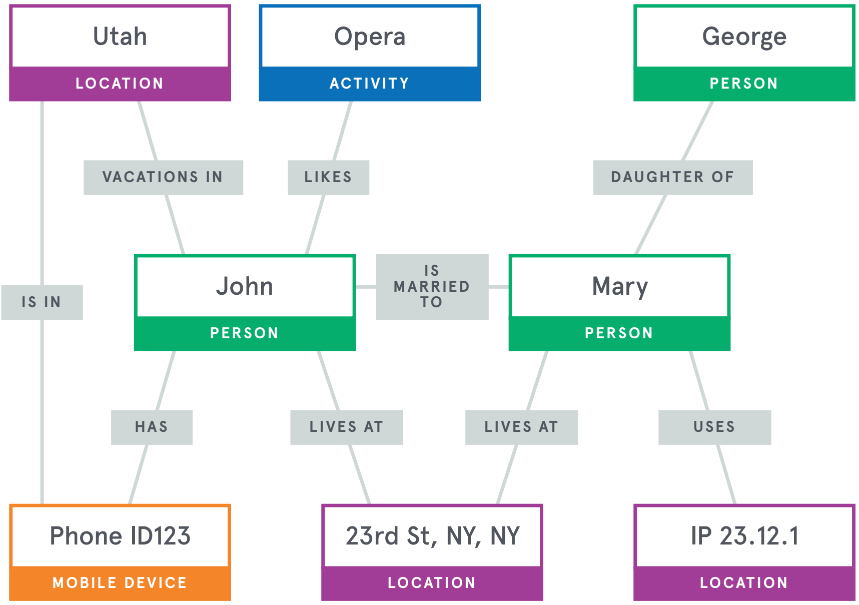


Figure 1: Identity Graph for John and Mary

[4] WorldWideWeb Consortium (W3C) Semantic Web https://www.w3.org/OWL/

THE TEMPORAL DIMENSION

In addition to static relationships, identity graphs can also incorporate a time dimension. While certain relationships are fixed, such as "Mary – daughter of – George," others such as John's location or Mary's location are continually changing. Figure 2 shows how identity inferences can be made through temporal relationships.

Figure 2 shows a mobile phone 123 at a particular geo-location (Lat A, Long B) at 2 p.m. The same device is then identified at a different geo-location (Lat C, Long D) at 2 a.m. If we assume the same person is in possession of the phone at both times, then we might infer that 23rd Street, New York, NY, is John's Work Address, and #1 Olive Street, New York, NY is his Home address.

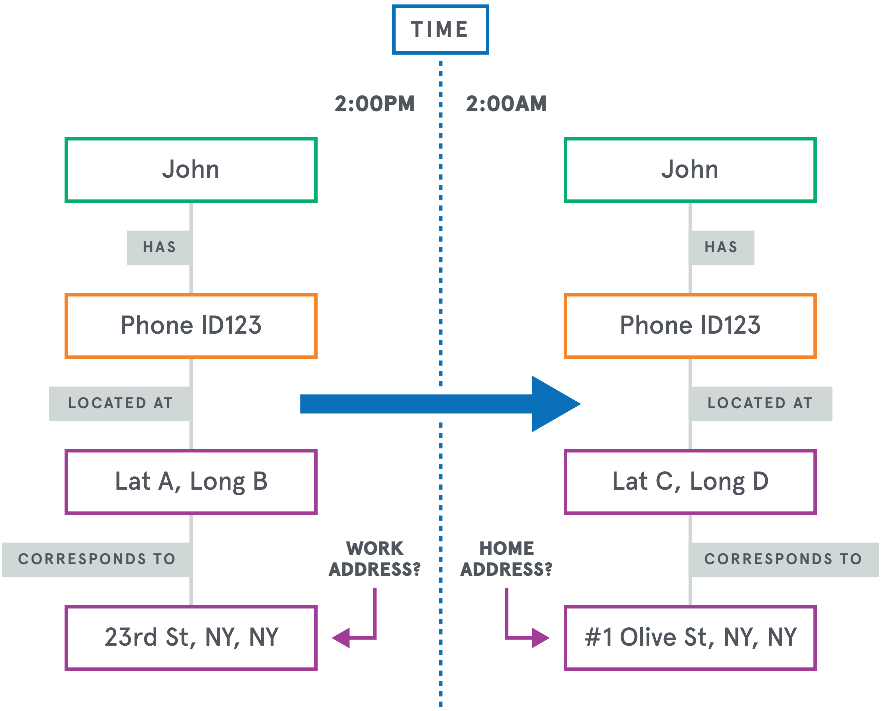


Figure 2: Inference from Temporal Dimension

The likelihood of this association is even stronger if the same pattern is observed over a period of weeks or months. Repetition over time is a critical factor in interpreting patterns and can substantially increase the accuracy of pattern inference. For example, from a probabilistic matching perspective, linking John Smith living at address X with a John Smith living at address Y is an unlikely inference because the name John Smith is widespread. On the other hand, if we observe the pattern John Smith and Mary Smith at address X and John Smith and Mary Smith at address Y, then the likelihood of linking both John Smith and Mary Smith to that address substantially increases. Even more so if we see the same pattern across three addresses X, Y, and Z.

ACCOMMODATING NEW RELATIONSHIPS

An essential feature of an identity graph system is the ability to easily enhance the model with new types of entities and new relationships. In this respect, we can learn a lot from systems designed to manage engineering and material assets such as parts and equipment. From the beginning, these systems have had to deal with an enormous array of entity types from bolts, screws, and ball bearings to electric motors, pumps, and generators. Because each part and assembly has its own unique set of characteristics, these systems were early adopters of a "graph view" of entity attributes. They learned early on that a significant aspect of identity management is the ability to easily create and manage new types of entities and connect them with existing entities.

Visualizing the relationships between identity fragments in a graph can open new opportunities. Finding connected components through transitivity (following a series of connected edges) can produce evidence for or against collecting fragments into the same identity. Graphs can also reveal natural groupings such as households, work cohorts, and social relationships.

No data modeler, regardless of experience or domain expertise, can anticipate all use cases for data. The current trend in system design has been to move away from the idea of creating a fixed model first and then transforming all incoming data into the model. Instead, system designers are looking for ways to incorporate more agility into systems, allowing them to accommodate change more easily. Graph-based systems and data lakes are part of this new paradigm in system design.

The current trend in system design has been to move away from the idea of creating a fixed model first and then transforming all incoming data into the model.

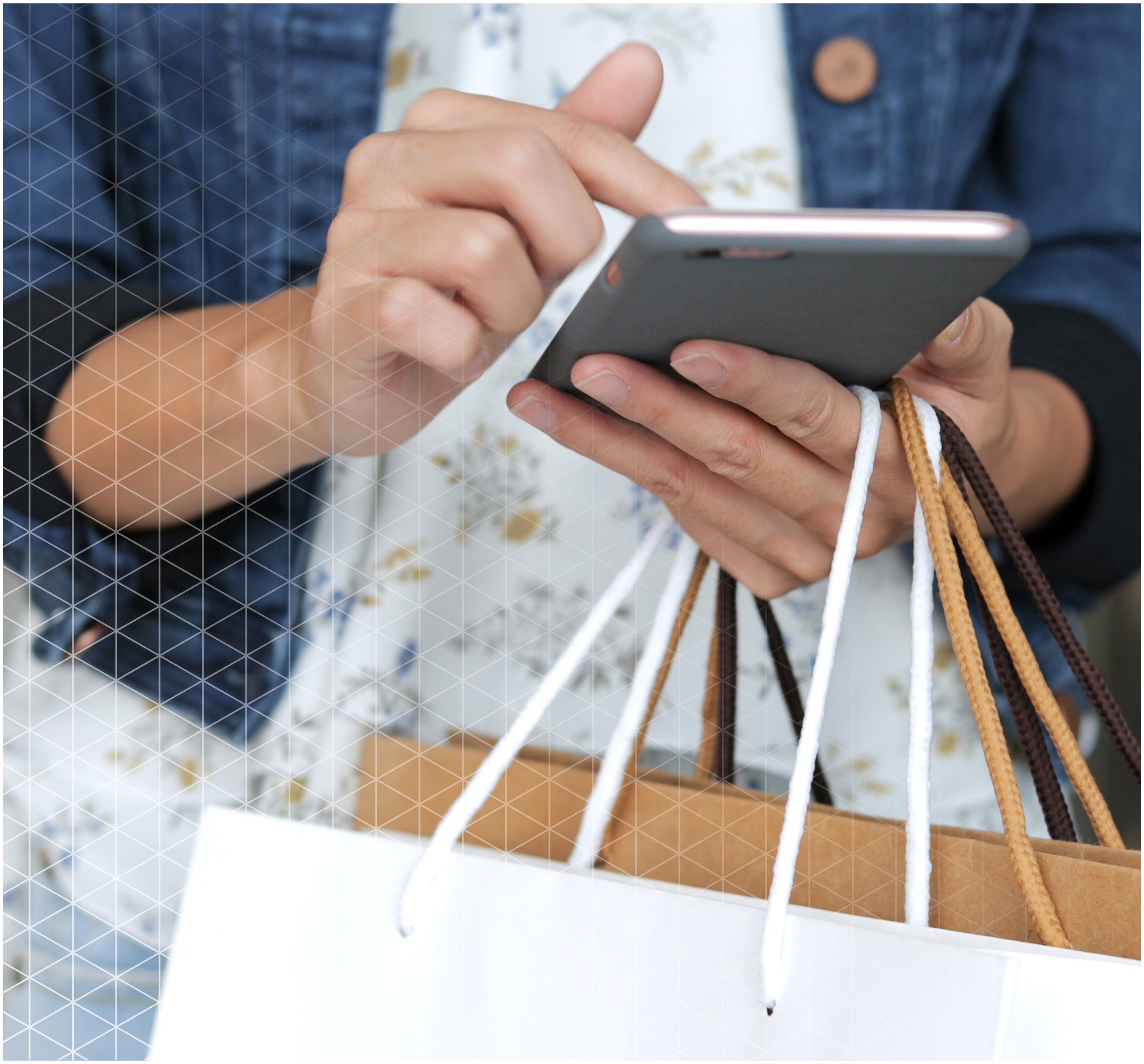
BRIDGING ONLINE AND OFFLINE AND THE PROMISE OF MACHINE LEARNING

Another vital feature of personalization is the ability to bridge the divide between the world of online, unstructured data and offline, structured data. As consumers conduct more and more of their social communication, shopping, and business transactions online, traditionally structured transactional information is providing fewer and fewer pieces of the customer preference mosaic. To fully understand customer preference, systems must continually ingest and scan vast volumes of unstructured information to try to identify and connect the relevant items into the identity graph.

Fortunately, recent advances in machine learning have suggested a path forward, but the unstructured world of social media still presents several challenges. Perhaps the greatest challenge is the fact that the best machine learning techniques are based on supervised learning. Supervised learning is a type of machine learning where the system must first be trained with a large number of examples inputs together with the desired outputs or decisions. This is as opposed to unsupervised learning, where a training step is not required. The type of data available for identity resolution is increasingly unsuited to supervised learning approaches.

Supervised learning has a long history in entity and identity resolution starting with the advent of probabilistic matching in the 1970s. The work of Canadian statisticians Ivan Fellegi and Sunter[5] is still the foundation for the current approach to probabilistic matching. However, their method for the calculation of probabilistic weights is based on a number of assumptions. One is that the entity references are structured, another is that the correct links between the references (at least for some large sample of the references) are already known. These are constraints we would like to move away from, as it is often not feasible to perform the necessary standardization or obtain sufficient training data to support supervised learning.

Much of the current research in entity resolution is directed at overcoming these types of constraints. New methods for word and phrase embedding and matrix comparison are showing promise for working directly with unstructured and heterogeneously structured references. New work on hybrid supervised-unsupervised learning is focusing on the technique of "bootstrapping" as well as active or reinforced learning. In a bootstrap approach, a partial truth about which references should be linked is produced by applying traditional unsupervised matching techniques such as deterministic matching. The partial truth is then used for training in a supervised model. When the new model is found to be wrong, the incorrect links are not only corrected in the output, they are also fed back into the system to retrain and improve the model, a process called "active learning."



[5] *A Theory for Record Linkage.* Journal of the American Statistical Association, Dec 1969.

THE FUTURE OF IDENTITY RESOLUTION

By embracing a more holistic approach, entity and identity resolution is evolving and adapting to the changing datasphere. It has moved from an exercise in writing database queries to true data science applications using statistical analysis and machine learning.

Many challenges still lie ahead in developing truly robust holistic identity resolution systems. Scalability has always been a challenge for entity and identity resolution even before the age of Big Data. Both defined in terms of pairwise comparisons which increase as the square of the number of records. Processing large numbers of entity references call for new iterative approaches to blocking and matching in order to scale in a distributed computing environment.

Concurrent with increasing scale is also increasing complexity driven by the need to recognize more types of entities, attributes, and relationships, to accommodate new dimensions of resolution, and the vast number of possible connections. Fortunately, the capabilities of machine learning and other artificial intelligence techniques are rapidly advancing. While its application to identity resolution is historically limited, research is ongoing to understand how these techniques can be effective in analyzing these large, complex networks and accurately resolve identities.

While much more research and development will be needed to create the next generation of holistic identity resolution systems, some short-term solutions are already emerging. Businesses are now able to buy an online identity graph for connecting cookies, MAIDs, and some email addresses from third-party vendors. These graphs may even combine some aspects of deterministic and probabilistic matching such as IP co-occurrence, then join the result to their offline data. However, holistic systems of the future should have the ability to simultaneously resolve identity using online and offline data as well as temporal and inter-identity relationships. In other words, the fully holistic approach combines deterministic and probabilistic linking decisions synchronously rather than sequentially.

At the same time, these more powerful identity resolution systems must continue to balance privacy, convenience, relevance, and security. Single point-solutions are no longer effective in addressing the wide range of digital impressions left behind by consumers across an ever-changing digital landscape. Only a holistic approach to identity resolution invoking a combination of many methods and tools will be able to support the level of personalized customer experience expected in modern e-commerce.

John R. Talburt, PhD
Director, Center for Advanced Research in Entity Resolution and Information Quality (ERIQ), University of Arkansas at Little Rock

Richard Y. Wang, PhD
Director, Chief Data Officer & Info Quality Program at Massachusetts Institute of Technology (MIT)

Dirk Beyer, PhD
Senior Vice President and General Manager, OneID and Data Science, Neustar

WP-MS-185274-08.12.2019