Advancing Customer Identity Resolution

The Neustar Federated Approach

Table of Contents

Executive Summary

Marketing technology players have, to date, developed three main approaches to solving the complexity of customer identity data resolution—each attempting to build a rich picture of the customer that is accurate and scalable.

Who Should Read This White Paper

- Sophisticated B2C marketers at enterprise-level brands
- Data science and technology leads at enterprise-level brands

Questions This White Paper Will Answer

- Why—contrary to popular belief—deterministic is not an inherently better or worse identity approach than probabilistic.
- What challenges and issues arise from merging deterministic and probabilistic identity into a simple, hybrid graph.
- How a new, Neustar 'federated' model resolves for leading approach shortcomings and produces significantly better results.

These approaches each succeed to some measure, but all are ultimately flawed, unreliable methods to resolving customer identity:

- The commonly used deterministic approach can view a customer's identity online and offline alike—but often breaks down in common, but complex real-world scenarios.
- The probabilistic approach uses pattern recognition to solve for those real-world complexities. However, it has no view into a customer's offline life.
- Hybrid identity attempts to capture the accuracies and cross-channel reach of both, by sequentially merging customer profiles from separate deterministic and probabilistic providers. But with no oversight of how those providers operate, complicated further by the challenge of blending data sets blindly, a simple approach will result in inaccuracy, data loss, and confusion.

Neustar set out to crack the identity code and tackle the inherent challenge in getting customer identity absolutely right.

We did so by harnessing competencies that set Neustar apart: extraordinarily high volumes of quality customer data, an extensive digital event-tracking network; specialized AI capabilities; and years spent maturing our capabilities in data science, thoroughly testing both deterministic and probabilistic approaches.

Armed with these differentiators, we have developed our breakthrough ***Neustar federated identity*** methodology: a new approach that builds identity clusters from the ground up, using only the validated linkages, seamlessly integrating deterministic and probabilistic sets of identity. The result is uniquely accurate and scalable, and provides a view of customers across platforms, devices, and into the real world.

In testing all approaches, we have discovered that:

- The pure deterministic approach offers the least coverage.
- The pure probabilistic approach offers no offline identity verification.
- The Neustar federated approach yields **10x the scale** of the purely deterministic approach.
- The Neustar federated approach yields **1.6x the accuracy** of the competing hybrid methodologies.

This white paper lays out where, and why, other solutions have failed and why cracking the identity code has proven so elusive. It also details how federated identity, above all other identity alternatives, enables marketers to take full advantage of the promise of data-driven customer engagement.

Customer Identity:

WHERE DATA HAS FAILED MARKETERS

Your sister comes to your house and borrows your laptop, using it to log on to a website she uses regularly. For the two of you, it is a common encounter. For brands seeking critical customer intelligence, it poses a hugely frustrating challenge. How can brands know to disregard the conflicting signal, and confidently know that the laptop is yours, not your sister's.

We're at the beginning of the people-based marketing era. With unprecedented data to identify the customer—across a growing number of addressable touchpoints—marketers can now evaluate countless interactions across the customer journey and target individuals with precision like never before. But to know the customer, you first must identify the customer.

Most providers, however, have yet to offer a meaningful, effective solution to customer identity. Let's be clear: The science of creating a coherent picture of people as they jump across addressable channels and devices, and interact via many different identifiers—from email addresses, to device IDs, to home addresses, and more—is exceptionally difficult. But identity providers must continue challenging themselves to innovate, since customer identity resolution is the critical foundation for marketing activities including:

- **Identifying and observing customer actions** and responses across the customer journey.
- **Measuring the impact** of every addressable touch across that journey.
- **Personalizing the experience,** including coordinating creative from one channel to the next; frequency capping across all addressable channels at once; and syncing in-store and call-center experiences with marketing engagements.
- **Building and engaging more effective audiences,** based on observations about how existing customers behave across the journey—coupled with demographic and psychographic data about who those best customers are and what they are looking to buy.

Machines that "See:" The Science of Entity Resolution

Identity resolution is a subset of ***entity resolution***—the field of AI that allows systems to distinguish one thing from the next. Formally, entity resolution seeks to determine whether or not different entity representations, such as cookies or device IDs, correspond to the same real-world entity, for example a person or a household. In real life terms, it enables a self-driving car to tell a standing pedestrian from an oncoming bus, or a photo application to tag a person's face in any photo, even as that person's expression changes.

Turning Single Identifiers Into Powerful Clusters of People, Places, and Things

Customer identity resolution matches each customer's many identifiers (or *signals*) into accurate *linkages*, and connects the linkages into larger *clusters*. These clusters are webs of identifying information that create a more complete picture—not just of people, but of places, and things as well:

People. Matching addressable "signals" like emails, phone numbers, and digital user IDs; it also means recognizing that many variations of an identifier all represent the same person.

- Ex: J Doe, Jon Doe, and Jonathan Doe are one person, not three.

Places. Matching a person, or multiple individuals, to a physical location, which could be a home or business address or an IP address matched to a geographic area.

- Ex: Jon Doe's email address is jdoe@website.com and he lives at 123 Main Street

Things. Matching devices and connected technologies

- Ex: The same person may own multiple addressable devices and use Bluetooth and Wi-Fi. These signals should also recognize that linkages for two different ads, one on a mobile device and another on an addressable TV, were sent to the same person.

Deterministic, Probabilistic, and Hybrid:

WHERE EACH APPROACH FALLS SHORT

We began with an example of a common, but tricky, identity problem: Your sister walks into your home and borrows your laptop to sign in to a website. The providers observing you and your sister's data are left to determine what part of that interaction accurately represents you (answer: the location and device), and what represents your sister (answer: the email address).

It is a valuable lens to view how each identity model operates, and a litmus test that shows where these approaches to identity resolution break down.

Deterministic: A Binary Approach That Struggles With Real-World Complexity

The commonly used deterministic approach is *not* likely to know the difference between you and your sister in the aforementioned example. It may (inaccurately) connect your desktop with your sister's browsing behavior, yielding the conclusion that the browsing session was yours and that your computer is your sister's. Such erroneous reasoning will almost certainly lead to problems in how subsequent behaviors and identifiers are interpreted.

The deterministic approach relies on *direct record-level observation of identity pairs*. In other words, it pairs identifiers by spotting instances where two identifiers have "interacted" in a clearly observable way. Here is how it works:

- The deterministic approach starts out with a customer-provided identifier—almost always an email address that the customer offers as login credentials to a website. The identity provider can then match that email to other identifiers through additional sources, including offline data.
- The identity provider can then look at the cookies connected to a customer's logged-in session and add that cookie to the customer identity profile.
- From there, the identity provider continues to connect those identifiers to others, creating linkages. For instance, if the customer's deterministic identifier, let's say an email address, is used against a certain cookie, then that device can be added to the linkage set. If it can connect a second home address or phone number to that device, it can add those identities to the profile as well.

The provider can then continue to create linkage clusters ad-infinitum.

The deterministic approach is essentially binary: for each potential linkage, it asks whether the data points to a connection or not. It can be highly accurate when the relationship between identifying sources is simple and straightforward.

This approach works when, for example, a person never shares his or her laptop with someone else. But that binary approach runs into trouble in a number of very common but complex real-world scenarios, as you can see in the issues below.

Keeping Up With Changing Customer Data

Deterministic clusters are, as noted, typically built using a customer-provided ID like an email address. But most people typically have multiple email addresses: a work email, a personal email or two, maybe one used specifically for online purchases to circumvent a barrage of marketing communications. The deterministic approach will create a customer profile using one email address, but to understand the whole person, their whole footprint and, ultimately, their identity, a bigger picture must prevail. Multiple email addresses must resolve to a single person. This same is true when looking at other deterministic signals like name, address, and phone number. As people change, so too does their identity, at least in terms of the data that represents them. They collect and change identifiers, which cause confusion when using purely deterministic data to manage and validate consumer profiles. Without a process in place to clean, merge, and maintain accurate offline linkages, identity data becomes out of date and invalid.

Inability to Decipher Multiple, Conflicting Connections

There are often many connections between identifiers (known technically as conflicting signals), such as the example of the sister borrowing a laptop. To get to the right identity answer, identity resolution must go beyond determining whether a connection simply exists to knowing which connections are meaningful, and which are not. From a brand's perspective, your sister's connection to your laptop is of little value; *your* connection to your laptop, however, can be extremely important.

Deterministic identity, however, struggles to recognize gradations of connection. Instead, with its binary either/or approach, it often decides that *all* existing connections are meaningful, thus leading to the incorrect deduction that your laptop belongs to your sister.

The deterministic approach also struggles with other common conflicting signal approaches which include:

- An email address used to log on to a website *might* be an instance of a customer using her own email; often, however, the user is borrowing a friend's login credentials.[1]
- Many devices that share Wi-Fi could indicate that those devices share an owner, belong to the same household, or are strangers all accessing the same airport hotspot.
- Telecoms rotate IPs across households at regular intervals, so devices using the same IP address may belong to the same household—*or they may not*.
- An Airbnb host sharing Wi-Fi with a different guest every weekend.
- Small business employees, and many family members, regularly use shared email addresses.

Does Not Search For "Missing" Information

The deterministic approach makes pairs out of readily-available information. However, the important information is often hidden from view—and the deterministic approach is not designed to take that "hidden" information into account. Two of the most serious examples of this problem stem from the very first step in the deterministic process: anchoring identity to a single email address.

Sees Only Part of the Picture

The deterministic approach builds identity off of one email address. The reality is that most people have multiple email accounts. That reliance on one email address alone looks at just one fraction of a customer's identity, without looking at other key parts of the customer's persona. But even when a customer's multiple email addresses are considered, there is a good chance the deterministic approach would view each email address as belonging to separate, distinct individuals.

Leaves Many Customers Unnoticed

If you log onto a website using your email address, a deterministic linkage has been made between your email and a cookie. A marketer can build on that initial linkage to build a deterministic identity profile. But what if you don't provide an email or any explicit piece of data identifying who you are?

Most of us will not think twice about providing an email to a utility company when paying a bill or to a bank when checking our balance, but when we are shopping or browsing online

we often will not provide—or even be asked to provide—a known identifier. This digital "window shopping" creates a blind spot for many marketers as they struggle to connect the dots across a customer journey, online to offline and offline to online. The result of this exclusive reliance on deterministic signals is that a large percentage of your customers to go unnoticed as they move across the web.

Deterministic Increases the Potential for Bad Data... and Bad Actors

The problems above are compounded by the fact that beyond just creating confusion and gaps, the deterministic approach is more susceptible to incorporating bad data—including data from malicious actors. Much of the bad data comes from a deterministic linkage supply chain that is hard to oversee, and from bots.

Commoditization of Data

Deterministic linkages are a commodity, and there is an economic ecosystem in which they are traded. Deterministic linkages are sourced from publishers that monetize their registrations and email/cookie pairs are sold as a way to generate incremental revenue beyond their advertising model.

Let's say Publisher A sells its deterministic data for a $10 CPM. They have an economic incentive to maximize the scale of their data set. In order to do that, they make affiliate deals downstream. These downstream players sell them their data for, say, a $6 CPM, making Publisher A a $4 CPM profit. The downstream partners also have an economic incentive to maximize the scale of their data, which can result in fraudulent activity. For example, some bad actors create fake email addresses matched with random cookies and then pass those fraudulent linkages along to publishers. The result is that publishers may unwittingly pass along those same fraudulent linkages on to data companies.

Neustar's own tests, using reliable first-party data sets and trusted third-party panel data, consistently find bad linkages riddled with inaccuracy. (These tests leverage data sets and models from our security, risk, and fraud detection offerings—which are crafted specifically to compare against other data to pinpoint inaccuracies and inconsistencies.)

Open to Bot Traffic

An estimated 52% of all 2016 web traffic came from bot traffic;[2] and in order for identity systems to know to ignore it, that system must be able to spot any non-human traffic.

Because the deterministic approach asks only if a connection exists between identifying data—and asks few, if any, qualifying connections about that data—it doesn't know to look for clues that indicate when data ought to be disregarded.

"Unfriendly" bots—designed to trick websites into thinking that they are human— are still more complex. For instance: If a bot steals a customer's email address, uses that email to log into a website, and a cookie is assigned to that login, how can a deterministic identity system know to ignore that information from its identity profile?

Bot-based fraud is not a minor issue, by any means. Recent estimates calculate that 35% of U.S. programmatic advertising activity was fraudulent in Q1 2017,[3] and that as much as 21.8% of all 2017 web traffic may have been driven by malicious bot activity.[4]

Probabilistic: High Digital Accuracy; Zero "Real-World" Visibility

Let us return to the example of your sister logging in to a site on your computer. The probabilistic approach will have a high likelihood of passing the litmus test, recognizing that your device is not your sister's, and that your sister's behavior on your device is not yours.

Probabilistic Spots Incorrect Information

By looking at patterns and linkage strength, the probabilistic approach can weed out many of the bad data and conflicting signal challenges that weigh down deterministic methods.

The probabilistic approach works by connecting identifiers through pattern recognition. For example, if records show that the same mobile and desktop cookie IDs visit a website from the same residential IP addresses, several nights weekly between 9PM and 10PM, then probabilistic methods may use that data to help determine that the mobile IDs, cookies, and IP address all belong to the same single user visiting from his or her home.

While the deterministic approach asks if two identifiers are connected, the probabilistic approach looks to patterns to answer the qualitative question of how connected the various identifiers are. For instance, a probabilistic approach may look at how often a device was used at a particular IP address, how often a set of devices shared a single Wi-Fi, or find other identifying patterns those devices within the same household have in common.

If a pattern emerges of those identifiers being connected, then the linkage is probably significant; if no pattern emerges, then it is probably incidental and can be ignored. The significance of each linkage is ranked on a linkage strength scale, from zero to one, where zero reflects no confidence in a linkage, and one implies maximum confidence.

The Benefits of Pattern-Seeking

By looking at patterns and the qualitative questions of linkage strength, the probabilistic approach can provide accurate insight in many situations when the deterministic approach breaks down.

Probabilistic Breaks Free of the Email Anchor

Unlike the deterministic approach, probabilistic customer identity is not reliant on a single anchoring email (or any other single identifier). Instead, it can use many sources at once to flesh out a customer identity—even when there is no email available.

- This means probabilistic can scale without relying on the often-missing single email identifier.
- Probabilistic also does not need to turn to problematic data sources to compensate for lost scale.
- It can build up a rich view of customer information based on a variety of sources at once—not just a sliver built out of one email at a time.

Probabilistic is Not Binary

The probabilistic model better allows for the uncertainty of the real world than the binary deterministic approach. In a sense, the shift from a deterministic to probabilistic approach parallels another shift in marketing analytics: from last-click measurement, which credits 100% of marketing business impact to the last touch before a conversion; to the more nuanced multi-touch attribution that looks at the full range of brand engagements and credits each touch with a fraction of the sale. Last click measurement links two data points that are relatively easy to observe—a final marketing touch, and a recorded conversion—and immediately assumes a 100% strength between the links, often wildly over-attributing credit to the last transaction. Multi-touch attribution takes advantage of developments in data modeling and machine learning to be able to recognize gradations in the linkages.

In both multi-touch attribution and probabilistic linkages, the shift to looking at the whole data picture-and not just the direct links-allows for a subtler, and often more accurate, understanding of the customer.

For instance, if a brand sees that two devices visit the brand's website via the same IP address then both approaches will recognize that the devices go together. But probabilistic approaches also allow for degrees of connection strength. A probabilistic approach might also look at other factors, such as how frequently cookies on one device match with cookies on the other device, how frequently both devices connect from that IP address, or how many other users connect from the same IP address. That additional information will help show if both devices might belong to the same person-a very strong link; or if they might be linked more loosely, such as devices owned by separate people in the same household or place of work; or if they're so weakly linked that they're only very distantly connected-for instance, they might be two strangers sharing the same corporate IP address

Drawback: No Window Offline

While digital pattern recognition is useful in building acquisition and targeting strategies to engage with customers online, it does not provide a tie to the offline world. Most marketers rely on an omnichannel marketing approach to deliver relevant experiences across a customer's journey as they move between the offline and online worlds. Without these offline identifiers marketers will only be able to reach their customers on their devices.

This is a serious drawback to the probabilistic approach in a world where close to 80% of shoppers still favor brick-and-mortar outlets for half or more of their shopping.[5] Statistics show that 75% of branded apparel purchases are made in-store,[6] and a full 67% of millennials head to brick-and-mortar locations to make purchases after first doing shopping research online.[7]

In Search of a More Intelligent Model

Neither deterministic nor probabilistic approaches are consistently correct or incorrect. Both produce a high percentage of accurate linkages, but deterministic is designed for a high degree of *accuracy under perfect conditions only*, while the probabilistic model better allows for the uncertainty of the real world.

Given their strengths and drawbacks, the ideal identity resolution methodology should take the best of both approaches, while avoiding the drawbacks of each. In theory, that should work. The practical reality is quite different.

Hybrid Identity: A Simple Approach

Simple hybrid identity is a sequential approach some identity providers use to compensate for both approaches' weaknesses, while attempting to capitalize on their respective advantages. It takes deterministic linkages from a deterministic identity vendor, probabilistic linkages from a probabilistic identity vendor, and then merges the two linkage sets together to form new, combined clusters.

But with this approach, if your sister used your laptop to log into a website, the outcome of this would still be unreliable, and may even assume you and your sister (who does not live with you) share a single, very large household cluster. Let us investigate why.

The problems emerge because of how the separate linkages are gathered: they are bought whole cloth from third-party providers. Simple hybrid identity does not know how its initial linkage sets are produced—and whether they are accurate. This is a similar issue to the deterministic data quality problem, outlined above (see "Deterministic Increases the Potential for Bad Data...and Bad Actors").

The lack of visibility into the third party "black box" opens the simple hybrid identity approach up to four separate problems:

1. **Poor Quality Raw Data:** Did the linkage provider create linkages off of accurate, high quality, up-to-date data? By simply accepting third-party linkages outright, simple hybrid providers are left unable to answer these and related questions into the quality of the "raw materials."
2. **Inaccurate Linkages:** Simple hybrid identity only makes sense if the linkages from both providers are accurate to begin with. But even when starting with the best raw data, every linkage provider will inevitably produce a certain number of inaccurate linkages. Plus, accuracy can vary quite a bit depending on how each methodology is applied. By implicitly assuming that every linkages is correct, the simple hybrid approach turns a blind eye to the inaccuracies it inevitably absorbs.

3. **Bad Linkages Taint the Good Ones:** Because linkage data is *linked*, one mistake in a cluster can impact the *entire* cluster—and the entire profile of the individual or household. This is a real risk when pushing together linkages from different sources, and potentially different levels of quality and accuracy.

   To see how that could happen, let us go back to our litmus test of your sister using your laptop. Presume the probabilistic linkages have correctly grouped your devices distinctly from your sister's. We will also say that the deterministic linkages, based on that single interaction, have *incorrectly* assumed that you and your sister share a laptop—and share all the other devices associated with your identifiers *and* hers. The result is the misperception of a single, massive, multi-device household.

   A single bad linkage in the mix has essentially "infected" the two accurate clusters—and has thus rendered not one, but two customers' identities inaccurate.

4. **Data Loss:** Matching different linkage sets is subtle work, and it is easy to miss crucial ways the different linkages should align. When you are dealing with two linkage sets from two completely disconnected sources, the inevitable data loss and missed connections become exceedingly more likely.

   To see this problem in action, let us construct another example: this time, a tablet at home, a laptop at work, and a phone you use in both locations. In theory, the phone can act as the "missing link" that connects all three of your devices—and possibly three locations—together. But the simple hybrid model may not be able to make that connection.

   Why is that? Let us assume a probabilistic set of linkages has connected the phone to the home tablet and a second set of deterministic linkages has connected the phone to the work laptop. In the process of connecting millions of data points across linkage sets, it is easy to overlook that single identifier—the mobile device—that could have connected both work and home.

In short, all three of the prior identity resolution approaches have had their degree of success; but also ultimately fall far short of the intended goal of accurately resolving customer identifiers—and creating one view of the customer.

HOW NEUSTAR FEDERATED IDENTITY RESOLUTION IMPROVES ACCURACY & SCALE

Neustar federated identity not only combines probabilistic linkages and deterministic ones but also considers and scrutinizes them synchronously (in parallel)—for the most accurate linkages and a full view of offline identity. This works toward the same end as simple hybrid identity, but federated identity takes an approach that is radically different, and vastly more effective. Rather than accepting linkages from third party sources blindly and sequentially, federated identity:

- Assesses and builds *all* linkages—deterministic and probabilistic alike—simultaneously and from the ground up.
- Tests all linkages for accuracy before incorporating them into the final linkage clusters.
- Merges the linkage sets into clusters that provide a highly accurate, complete view of the customer.

With complete control and oversight over the linkages, we can:

- **Maintain Quality Controls.** We have created the linkage, and so we understand the information that has gone into it—and in particular, we know that we do not turn to questionable data sources.
- **Ensure Accuracy of Each Linkage.** Across every cluster it creates, the Neustar federated approach runs deterministic and probabilistic linkages side-by-side and validates its models against known data sets from first-party login data and trusted third-party panels. With those known data sets as a guide, it uses only those linkages—from either approach—that have the highest accuracy scores. The result is that all linkages are highly accurate—and we weed out any "rotten" linkages that can spoil the bunch.

Federated: to unite separate data sources into a more powerful, single whole.

- **Match Linkage Sets From the Start.** Creating linkages from the ground up lets us begin to match each deterministic cluster with its parallel probabilistic one—right from the moment of inception. This vastly reduces the problems of data loss and missed connections that the simple hybrid approach suffers from.
- **Uniquely Scale.** The federated approach can combine information from *both* approaches to uncover linkages that neither approach discovers on its own. For instance, if a deterministic linkage connects an email address to a desktop cookie, and a probabilistic linkage connects that cookie to a device ID, the Neustar federated approach can connect linkage types to combine email, cookie, and device ID together.

The Neustar federated approach is likely to know that the laptop your sister used in your household is not hers. By evaluating the deterministic and probabilistic signals simultaneously, any linkages found between your household laptop and your sister's login credentials will be discarded. Rather than a snapshot in time, Neustar looks at the history of linkages within your household to provide a more accurate record of identity rather than a record of those individuals who periodically visit your home or use your devices.

With this new and improved approach, Neustar believes it has cracked the code on identity resolution.

How Neustar Federated Identity Works

The Neustar federated model builds its identity profiles through a four-stage, synchronous process (*not* a sequential hybrid):

Data Gathering

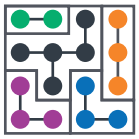
Neustar assembles a pool of hundreds of billions of identity data points across online and offline sources—ranging from existing customer files to cookie data, device IDs, and records of "events" like ad impressions and device pings.

Hypothesis Creation

From the "raw signals" in the first phase, the system carves out possible features that may represent a single digital device, person, or household and possible linkages between the identifiers. The system might arrange multiple cookies that likely all stem from the same mobile tablet, or multiple variations on a name (like J Doe, Jon Doe, and Jonathan Doe) that may each resolve to the same individual.

Linkage

The machine learning evaluates these hypotheses and decides which data points should be linked, and which should not. The result are "nodes" of connected data signals. (At this stage it is still unclear how strongly each linkage in the node should connect.)



Cluster Formation

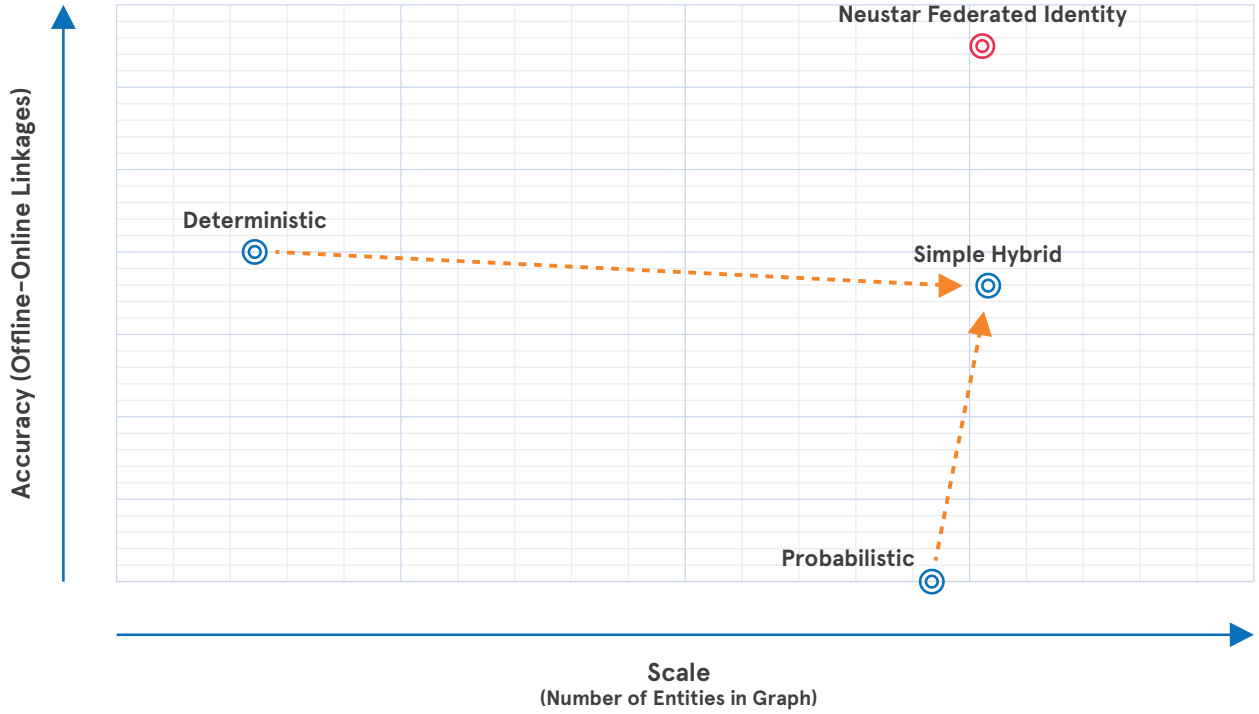
This last stage in the process looks at the strength between the linkages—to determine which identifiers resolve to the same person, place, or connected object (i.e. those that have a very strong connection to each other); which belong to the same household or institution; and which are connected but not in any meaningful way (as in the case of strangers sharing the same airport Wi-Fi). Machine learning arranges the meaningful linkages together into clusters.

Significantly More Accurate Than the Alternatives

How much more accurate is the Neustar federated model than the others? We took a statistically significant set of known linkages, broke them down into their raw identifier data, and ran them through each of the four approaches.

- We scored the results based on how accurate the models were at connecting offline and online data, and the "coverage"—how much of the overall linkages each model was able to capture.
- Consistent with the above rationale:
  - The pure deterministic approach offered the least coverage.
  - The pure probabilistic approach offered the weakest offline-online accuracy.
  - The federated approach **provides 10x the scale** of the purely deterministic approach.
  - The federated approach **provides 1.6x the accuracy** of a competing simple hybrid methodology.

CUSTOMER IDENTITY RESOLUTION APPROACHES



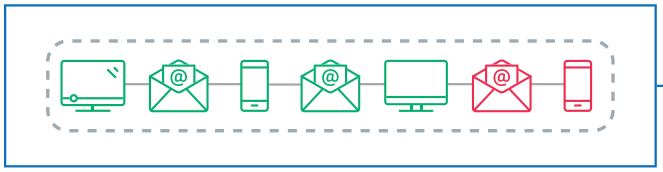
Building the Right Clusters

Simple Hybrid Graph

A simple, hybrid identity resolution approach promises the scale of probabilistic and the accuracy of deterministic, but fail to integrate the two methodologies successfully. Running the two models side by side is an important first step, but the ability to combine each methodology's best results into one, more accurate identity graph, is where the Neustar federated model is far more successful than simple hybrid methods.

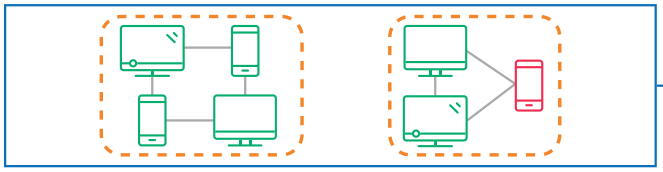
Take a look at the way a simple hybrid graph gets formed, in contrast to the Neustar federated graph featured on the next page, using the example of your sister (whose use of your household computer is denoted in red).

GRAPH 1



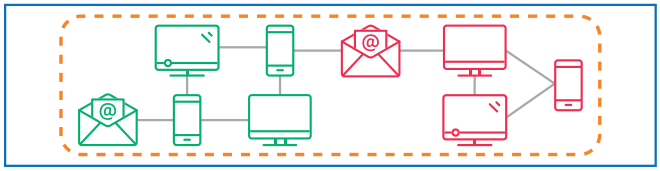
Graph 1 contributes a set of deterministically linked clusters identified within your household including your sister's logged session data, denoted in red.

GRAPH 2



Graph 2 contributes a set of probabilistic clusters associated with those deterministic nodes.

SIMPLE HYBRID GRAPH



One giant, multi-household cluster results from merging deterministic graph connections with probabilistic graph clusters.

A deterministic view (*Graph 1*) of your household would not distinguish your devices and login data from your sister's.

Using the probabilistic graph (*Graph 2*) to add scale to the deterministic clusters without evaluating the strength of those deterministic linkages can lead to problematic results.

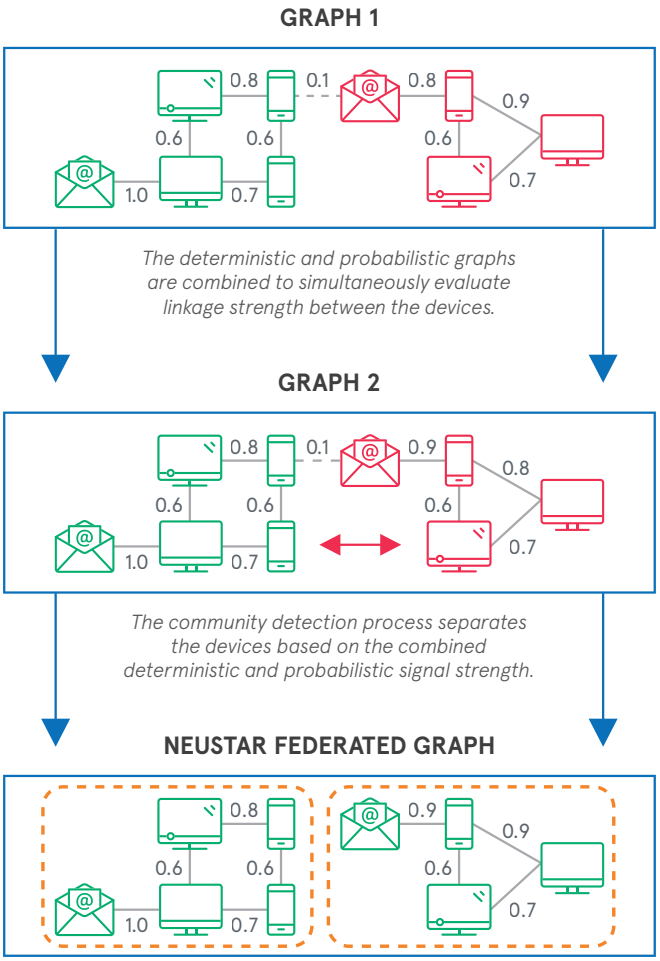
When the deterministic graph—which includes your sister's logged sessions—is then combined with a probabilistic graph (*Simple Hybrid Graph*) containing all of your sister and your household's associated devices and signals, the result is one, large multi-family cluster that incorporates your sister's household into your own.

From a marketing standpoint, such unreliable identity resolution methods can result in an entirely faulty identity infrastructure, which then trickles down into poor customer experiences and misspent media dollars.

Neustar Federated Graph

If instead, identity linkage strength were evaluated before all deterministic signals are accepted as meaningful, a different story would emerge.

In this federated identity resolution example, the process, known as "community detection," is used to build accurate clusters with reduced noise.



Two clusters are formed that combine the strengths of the probabilistic and deterministic data graphs separating your sister's household cluster from your household cluster.

As depicted in the first two graphs, the community detection process uses probabilistic clusters to evaluate the strength of associated deterministic linkages. Because the signals from your sister's devices and logged sessions are not statistically significant compare to the rest of the clustered signals within your household they can be ignored, reducing noise and increasing accuracy.

As the Neustar federated graph shows, your sister's logged web sessions and all her associated probabilistic identity data is identified as a separate household and is removed using community detection. The resulting clusters provide an accurate view of your household and—as resolved in green in the Neustar federated graph—your sister's separate household.

Why the Neustar Federated Identity Approach is Superior

Federated identity is a vastly complicated and sophisticated approach that requires enormous data and computing resources—and one that no other identity provider has yet to undertake.

Federated identity requires:

Vast data at the ready.

An enormous volume of data is important for all identity work—and is particularly critical for being able to make deterministic linkages, *without* needing to turn to questionable third-party data sources to supplement missing information. Neustar works with hundreds of billions of records on device IDs, cookies, locations, individual adults, and households.

A significant baseline of authoritative data

To check which links are accurate, you need a very significant sample size of authoritative information to check against. Neustar has a formidable portion of the hundreds of billions of records mentioned above included in our data sets and models from our security, risk, and fraud detection offerings. These are crafted specifically to help pinpoint inaccuracies and inconsistencies in other data sets.

Massive computing power, particularly in AI

The federated approach essentially calls for doing the work of not one, but multiple identity providers: acting as a deterministic provider, a probabilistic provider, and a third operator that tests all linkages for accuracy and connects the linkages together.

- This takes an enormous amount of computing power. In particular, it takes extremely sophisticated machine learning to:
  - Find patterns in the probabilistic signal to calculate linkage strengths
  - Determine the real accuracy of the deterministic linkages
  - Intelligently combine linkages to create accurate clusters
- Neustar AI capabilities include:
  - A core data set of known, first-party information, including offline and online identifiers, and data from trusted panel providers
  - Massive scalable machine learning resources.
  - AI enhanced by the vast data pool Neustar possesses

Experience in all models

Creating effective deterministic and probabilistic linkages in a synchronous (side-by-side) method requires first having mastered both approaches individually. Neustar has years of critical experience with both models in this regard.

Federated identity is a vastly complicated and sophisticated approach that requires enormous data and computing resources—and one that no other identity provider has yet to undertake.

Conclusion

Accurate identity resolution is fundamental to marketing. And yet, while the deterministic, probabilistic, and simple hybrid approaches all achieve a degree of success in providing accurate, scalable customer identity, **our tests reveal that they each fall short in important ways.**

Neustar federated identity, on the other hand, yields uniquely superior accuracy without sacrificing scale.

This ability to "connect the dots"—to succeed in the new science of customer identity management—is as complex as it is crucial. Done right, identity management helps foster impactful targeting, custom engagement, and precision measurement throughout the customer journey. In other words, it enables true people-based, customer-centric marketing. Done wrong, it lays the foundation for poor measurement; personalization gone awry; and vast sums in wasted data and advertising spend.

In short: there's a lot riding on the ability to coordinate billions of data points across millions of customers so we believe that no one methodology can get you there successfully. You need to look across all available options for connecting and linking customer data to build something accurate, something sustainable. The ideal identity resolution methodology would take the best of both approaches, while avoiding the drawbacks of each.

Appendix

1 This case is particularly common: For one indicative statistic: some 21% of 18 to 24 year olds streaming viewers have been found to share video streaming passwords with people who do not live with them.

2 Source: The Atlantic, 2017: The Internet Is Mostly Bots

3 Source: Digiday, 2018: Global State of Ad Fraud

4 Source: Info-security.com, 2018: Bad Bots Make Up a Fifth of All Web Traffic

5 Source: NRF Consumer View Fall 2017

6 Source: Oliverwyman.com, 2018: Infographic: Understanding the Apparel Shopper Journey

7 Source: Retaildive.com, 2017: Most millennials, especially younger ones, still prefer stores

About Neustar.

Neustar, Inc. is a leading global information services provider driving the connected world forward with responsible identity resolution. As a company built on a foundation of Privacy by Design, Neustar is depended upon by the world's largest corporations to help grow, guard and guide their businesses with the most complete understanding of how to connect people, places and things. Neustar's unique, accurate and real-time identity system, continuously corroborated through billions of transactions, empowers critical decisions across our clients' enterprise needs.

More information is available at

www.home.neustar

WP-MS-148085-02.11.2019