TAPAD
Unify life across devices

Measuring Cross-Device, The Methodology

As the first company to crack-the-code on cross-screen, Tapad Data Scientists are asked to explain the power of our cross-screen technology on a near-daily basis. This is their overview of our methodology and how we make it all work.

WHAT PROBLEMS ARE BEING ADDRESSED:

Stating the obvious, the ultimate goals in digital marketing are to run campaigns that work -- that is: increase transactions, brand awareness, whatever the metric - and to do so as efficiently as possible.

Reaching these goals requires two things:

1. Reaching the right people, on the right device with the right brand message (accuracy)
2. Connecting with enough people to positively impact the brand (scale)

Getting the right combination of scale and accuracy is no easy task; and right now, it's probably the most discussed topic in the cross-device advertising space.

Before we dig in, here are some definitions crucial to understanding how privacy-safe accuracy and scale can be achieved in cross-device marketing.

Deterministic Data Approach: Identifiable data (logins, emails, mailing addresses, e-commerce) is used to determine that a person is the same on one device as on another. Deterministic is also called first-party data, meaning that the data came from a company that collected the data itself -- it is known log-in data. For example, if someone signs into their Amazon account on a phone and then on a computer, they are identified as the same user.

Deterministic data provides extremely strong indicators, but given the potential for someone to log-in to say, Amazon, from someone else's computer, it cannot be considered 100% accurate -- in fact, no approach is. A good example is Comscore's recent validation of AOL's deterministic data set which landed at 93%.

Other questions that arise in deterministic data use include privacy, limited log-ins on mobile devices, and a hesitance for marketers to tie themselves to any single first-party data provider (i.e., Facebook, Google) - this is the "walled garden" that is also oft-discussed.

Probabilistic Data Approach: Non-identifiable data points are aggregated to determine the strongest probability that devices are related. Here, algorithms look for patterns such as device proximity (if devices are always seen with one another via internet connection then they're probably related), browsing patterns (looking at the types of sites and content that are consumed on devices) and time-based clues (analyzing when devices wake

up and sleep can help build the case for their being connected to the same user).

The value of probabilistic modeling is the ability to scale. Because the model doesn't rely on people logging in to a common platform (like social media or e-mail), this approach allows for a much larger pool of devices. This is important as new devices such as internet connected appliances, cars or wearables arise. These new devices probably will not call for users to log-in/ give identifiable information. For example, it's not likely users will log-in to Twitter on an internet-connected refrigerator. While probabilistic is excellent for scale, on its own, it is harder to prove accuracy.

Combined Data Approach - This uses a privacy-safe deterministic data set when available. In this case, the deterministic data set is also used to verify the accuracy of the probabilistic data.

Tapad uses the Combined Data approach. To appreciate the value of this methodology, understanding how accuracy and scale are achieved is paramount. Here goes.

THE RECIPE FOR ACCURACY

The way you arrive at accuracy involves many moving parts, particularly in a combined data approach.

To get started, here are the definitions you need to know:

True Negative: A correct prediction that two devices are NOT related.

True Positive: A correct prediction that two devices ARE related.

False Negative: An incorrect prediction that two devices ARE NOT related.

False Positive: An incorrect prediction that two devices ARE related.

True negatives and true positives are pretty straightforward. If the probabilistic algorithm shows that two devices are not connected, and the truth set knows they are not connected, then the outcome is a true negative (a correct prediction). A true positive is the same concept-- a prediction that two devices are connected, which is confirmed by the truth set.

False positives and false negatives are more complicated, but very important when looking for the combination of scale and accuracy. Both false positives and false negatives indicate errors in the system, but one is far more detrimental to the success of an advertising campaign.

Less detrimental is a false negative. This occurs when devices do not appear to be related even though they are both owned by the same person. So, 'User A' might have a laptop, smartphone and a connected TV, but only the laptop and smartphone are showing up as related devices. This isn't ideal for an advertiser, but the consequence is small. The user will still see ads, just not across all of their devices.

A false positive has a greater impact. This occurs when two devices appear to be related, even though they don't actually belong to the same user. When data and behavior information about multiple people are erroneously combined into a single user profile, the view of the consumer is skewed and the campaign is less effective.
Just as it is easy to augment a device data set to achieve scale, it is easy to shrink it to achieve accuracy. Probabilistic methods provide various levels of confidence that multiple devices are related. Connections with low levels of confidence can easily be removed for the purposes of testing. This would increase the accuracy and reduce scale.

Maintaining a low false positive rate while also having a low false negative rate AND scale is optimal. This combination is a strong indicator that the Device Graph in question was neither artificially augmented nor scrubbed.

True or False: True is Not Always False, False is Not Always True

It is easy to assume first-party/deterministic data sets will be 100% accurate. In fact, it brings complexities of its own. An example:

- User A logs into Amazon on a friend's computer. The friend gets the laptop back and clears the cookies so that all of User A's information is removed.
- The probabilistic data is cross-checked against a first-party data set -- and it shows User A is NOT related to this smartphone.
- The truth-set still cites a false negative.

In other words, in this case, the truth set was wrong.

There are several reasons this happens.

Expired device identifiers are a big culprit. Cookies and/or device IDs expire and not every data supplier removes them immediately. Since data cleansing happens at different intervals - varying company by company - not every truth set is going to say the same thing.

Here's how that can play out, going back to our friend, User A, who logged onto Amazon on a friend's phone:

- The cookie on the smartphone was cleared, but the truth set provider didn't reset their data so the connection with User A was never removed.
- The smartphone cookie expired after a long period of not being used, but because expired cookies were not cleared from the truth set, the connection with User A was never removed

Even though the probabilistic prediction was right, it still appeared to be a false negative. One false negative might not seem like a big deal, but if User A reset his cookie 25 times over the course of a year and the provider of the truth set never did a data cleanse -- the number of false negatives would get very high.

Are there ways to avoid this? Not entirely, but improvements can be made.

Let's start with a high-quality truth set. With a quality data set, one that is frequently cleansed, the false negatives are more likely to be issues within the original Device Graph, and the Graph can be adjusted accordingly. Starting with a poor quality truth set will make it virtually impossible to address -- much less identify -- flaws in the probabilistic model. Additionally, Tapad uses multiple, quality truth sets to verify our probabilistic outcomes. This culls out a high percentage of inherent flaws in any one data set (expired device identifiers, e.g.).

External validation is critical too - the publicly stated Nielsen confirmation of Tapad's accuracy is well-known (91.2%), and Tapad consistently runs private tests with some of the world's largest first-party data owners. Results in our continuous private testing consistently average in the mid-90s.

SCALE: The Fine Print

Of course scale matters, and it's pretty clear why. But how you get there matters too, and there are many twists and turns in a combined data approach.

Anyone can claim they have scale. But how do they prove it?

First, there are tools that only the company who built the Device Graph sees. Tapad, for example, runs a weekly report to track our match rate with all of our top suppliers. The match rate is the number of devices with a demonstrated relationship to additional devices and appear in both data sets - it is the data we have in common.

In match rates, it is all about the recency and relevance of data. Every truth set will contain data that does not align with the probabilistic data, such as:

- Expired identifiers (as explained above, this depends on the frequency of the data cleanse). The Tapad Device Graph constantly refreshes, removing all inactive devices and giving us an extremely clean and relevant data set.
- Tracking data. Once a user has opted-out of targeted advertising, Tapad no longer sees them.
- Browsers that block cookies (e.g., Safari on iPhone). If an iPhone user is logged into an app or is using a browser that does not block cookies, they will appear in the Tapad Device Graph.
- PII, such as user login data. Tapad is blind to all PII.

Companies will also look for validation against a data set outside of their supplier ecosystem -- again, starting with match rate. A low match rate against a company with a massive first-party data set is an indicator there might not be a great degree of scale overall.

Bottom Line: Why is a Combined Data Approach Best?

It comes down to two primary factors -- quality and quantity. In a probabilistic model, the accuracy of algorithms and the scale of data determines efficacy. By leveraging probabilistic

and deterministic, Tapad employs a combined approach to achieve optimal scale and privacy-safe accuracy.

SIDEBAR: Tapad's Combined Data Approach

The Device Graph -- which operates the longest running cross-device algorithm in the market today -- interprets trillions of data points every month. The primary job of the Device Graph is to determine if different devices are related (i.e., are owned by the same person).

Here's how it works.

First, our algorithm mines device data for connection clues. It then analyzes those clues to make a prediction about which devices belong to a single user. If the algorithm determines there is enough data to connect the devices, a cluster is formed in the Graph around a single, anonymous user.

At this stage, if something doesn't add up between the two, the Data Science team goes back to work to determine where the model was flawed, and they fix it. Push play.

The result, what we call a combined data approach, gives Tapad's clients all the benefits of the scale of probabilistic with the accuracy of deterministic.