The Thread

There's something happening here,
What it is ain't exactly clear
Buffalo Springfield 1967

On June 28th, Bill Harvey published an article in Media Village entitled: Cross-Media Duplication Must Be Rigorously and Empirically Determined. Bill emailed excerpts of that article on June 29th which snowballed into a thread joined by several of the industry's most senior media researchers.

PART I

It starts when you're always afraid,

It started with Bill's assertion in a note to Josh Chasin,

> Whereas Project Blueprint found that duplication among media is definitely not random, the WFA/ANA blueprint allows for a concept of Virtual IDs which will tend to produce results that are very similar to random probability.

In the June 28th Media Village article[1], Bill cited an analysis performed on Nielsen One data comparing Nielsen's measured duplication with that generated through a random duplication algorithm. Unsurprisingly, Bill notes significant differences between the Nielsen estimates of duplication and the random duplication estimates with little consistency between those differences.

> It is worth the read to understand the differences, but it is in my view a little less helter skelter. In 9 out of 10 cases of TV/CTV duplication, random duplication was the same order of magnitude but higher. In 9 out 10 cases of mobile/CTV duplication, random duplication was lower but of the same order of magnitude. Think about it, would we expect mobile and CTV to have more duplication than random and TV and CTV have less duplication than random?

There's battle lines being drawn,

From Josh to Bill,

> Hey, Bill. While I wholeheartedly agree as regards to (the accuracy) of random duplication, I think you are unfairly criticizing

[1] https://www.mediavillage.com/article/cross-media-duplication-must-be-rigorously-empirically-determined/

the VID construct. … this model calls for a cross screen panel as a 'source of truth' … for overlap and duplication, so that the VID model is trained on and replicates actual, empirical, observed duplication.

And while Josh found the "source of truth" concept a little shaky, he conferred with Michael Vinson,

So my understanding is – Michael, ….. that if executed properly, duplication in a VID construct should replicate that of the panel 'truth set'.

Michael's return,

In principle yes. The idea of the VID is everything you know about media consumption is represented. Or to put it differently, when done properly you won't be able to find a contradiction between the VID results and the estimates that went into them. Testing this assertion was the very first thing ComScore did in the POC work.

But then,

Separately, can we please not refer to a panel as a "truth set"? At best a panel is another source of limited intelligence.

Nobody's right if everybody's wrong,

A quick response from Joan FitzGerald,

I'm skeptical that we have moved past "panel as Truthset", since this concept is fundamental to how the ANA defines good measurement.

Bill cites the CIMM Sequent Janus study that refers to panels as benchmarks, as a more neutral term.

And from Alice Sylvester,

I'm concerned that shorthand statements like this (VIDS generate random duplication), will take hold and we will be undoing that notion for a while. If I am wrong, and VIDs= random probability, then we are all wasting our time.

PART II

What is a VID anyway?

The following link takes you to an ARF Cross-Platform Council Studio on the architecture of VID's. While labeled a layman's guide, it is most effective at identifying the complexity of VID architecture. https://youtu.be/8qBWwnW9G6g

Josh offer's a simpler illustration,

> the VID construct calls for what I call "faux" or "synthetic" IDs, representative of the entire universe. Let's imagine a world where there are only four media companies—Google, Facebook, Paramount, and NBCU. …With VIDs, there will be a VID we can think of as representing me. It might have my Google usage from Google, but some other handsome dude's Facebook usage data, and still another insanely witty guy's Paramount data, and someone else's NBCU data. In that sense, the VIDs are synthetic.

I like the simplicity of this example, but have one elucidation – the VID is a virtual person, no longer Josh. In a country with 260 mm persons, there will be 260 mm virtual people whose data are modeled off a combination of digital and panel data in a way that reproduces the duplication of the "Benchmark Panel". At the same time, noise is introduced to ensure that a virtual person cannot be mapped to a real person.

Paranoia Strikes Deep,
Into your life it will creep

There remains skepticism that the benchmark duplication can actually be reproduced for 260 million persons across more than 2 platforms:

From Joel Rubinson,

> There will never be an integration solution because of walled gardens and privacy concerns. You will need to model the covariance and use a consistent approach to simulate exposure records and calculate reach. In the data set I have, the correlations got into the 80% range between certain pairs of tactics and then I had negative correlations between those tactics and others.

Though Leslie Wood indicated that NCS had integrations with multiple walled gardens.

But from Mainak Mazumdar on two being magic (speaking of Nielsen's analysis),

> The noise in the hash increases (to protect privacy) as we increase the number of media entities. Adding a second media entity can reduce accuracy from 90% to 75%.

Adding a third, reduced the accuracy of duplication to 50%.

Other loose threads (but highly visible) in what we too harshly dropped into the paranoia section:

Several from Tony Jarvis's journal, for which we offer a few lightning strikes,

When is close enough, not close enough? (per Alice)
Can we produce a truthy-enough truth set? (per Jonathon and Josh)
Frequency is crabgrass. (per Erwin Ephron)
Will it work with print, OOH, Audio, etc.?
Whatever it started out as, it had become clear that it had become Facebook/Google/TV unduplicated. Two publishers and a medium.

The fourth strike pretty clearly will creep into the life of a media planner if as Mainak suggests, accuracy deteriorates to 50% for the third+ media, audio, print, ooh, ...

Getting so much resistance from behind,

Several comments on the thread followed Tony's fifth strike, the walled garden origins of the ANA CMM design. Several threaders noted the justifiable skepticism that the VAB and broadcasters might have of CMM as a "Trojan horse" but Josh who knows the authors of the design very well, was highly complimentary of their smarts and ability to design a workable system.

And finally,
For What Its Worth,

Over the course of the last week, over 13 industry leaders responded to this spontaneous discussion of one of the most important issues in the industry. The threads (there are two to follow) are over 40 pages long and the discussion so rich, we felt we had to memorialize it in a readable chunk.

Apologize to any we missed or to a poor correlation between contributions and attributions in this summary. As I write this, the night before publication, there are still emails on the subject coming in, and so, I am sure this will not be the last you will see on the topic.

This was fun. Can't be planned. A sign of a vibrant intellectual industry. Let's hope it happens again.

Best