Protecting Content with Enhanced Gini Entropy Analysis

A technical paper prepared for presentation at SCTE TechExpo24

Jeffrey E. Calkins
Lead Data Scientist
Charter Communications
jeffrey.calkins1@charter.com

Kei Foo
Senior Director
Charter Communications
kei.foo@charter.com

Srilal Weera
Principal Engineer
Charter Communications
srilal.weera@charter.com

Vipul Patel
Vice President
Charter Communications
vipul.patel@charter.com

Table of Contents

List of Figures

1. Introduction

Digital Rights Management (DRM) is a component of modern content security mechanisms. Its primary purpose is to mitigate content theft. Despite the successes, quelling content abuse has been an ongoing battle. The exploits could range from unauthorized password sharing and stolen credentials to automated bots masquerading as humans. Automated bot activities also add unnecessary load to IP video delivery systems, consuming capacity that could have been used to serve legitimate customers. Current fraud analytics, however, are generally based on aggregated trends and are not sufficiently granular. Automating traditional entropy-based methods to track millions of devices and transactions is also computationally expensive. In this paper, a novel approach to address these issues is described.

In MSO networks, the headend collects stream license request and viewership data from a multitude of customer devices. Over time, such data exhibits certain patterns due to the differences in individual viewing behaviors. A typical user's channel change behavior (such as via a TV remote), generally follows a regular pattern. Each such sequence also has an associated 'entropy' value, which is a measure of the 'diversity' of the channel change sequence. Quantifying the diversity would enable us to draw inferences on user characteristics and system conditions. We have analyzed the resulting probability distributions and were able to derive actionable outcomes ranging from detecting credential sharing/stealing and system anomalies to advertising opportunities. The data presented is generic and normalized to preserve anonymity.

2. Digital Rights Management (DRM) Overview

As content distributors continue to seek efficient ways to identify potential system abuse and fraud in IP video, logged data from IP video delivery systems is used to create metrics, defining the norms and identifying trends that are outside of the norm. To demonstrate how entropy identifies bots and non-human abuse in DRM license requests, below is a high-level architecture of IP video and its workflow at video playback.

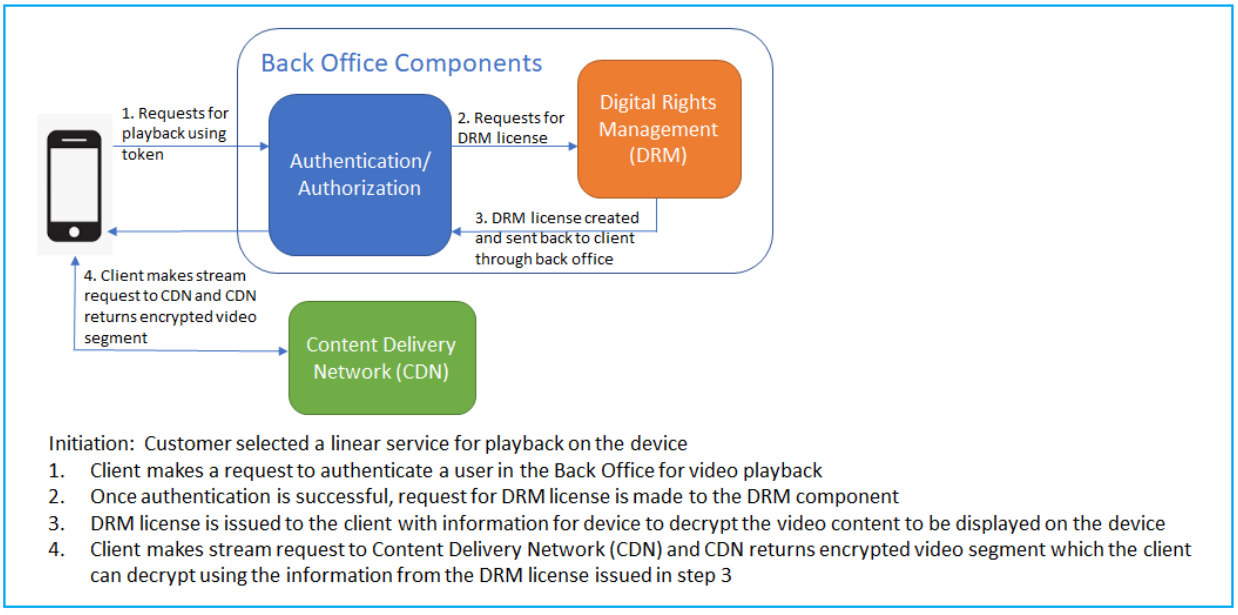

Figure 1 – DRM Workflow

Once the user is authenticated and authorized to playback video content on a device, the authentication module will pass the DRM license request to the DRM platform. Because a content protection mechanism DRM is implemented to keep video content secure through encryption, a DRM license containing decryption information must be provided to the client device for successful video playback. The DRM platform creates the DRM license and sends it back to the client through the back office. The client then makes a stream request to the CDN for the video segment and the client uses the information provided in the DRM license to decrypt the video and display on the device. To properly secure IP video content, each content has its own encryption. This means as a user performs a channel change, a new DRM license is needed. In addition, DRM licenses have expiration times necessitating new license requests. A valid token request is made by the client prior to making a DRM license request. In the current workflow configuration, the number of token requests and DRM license requests have similar values.

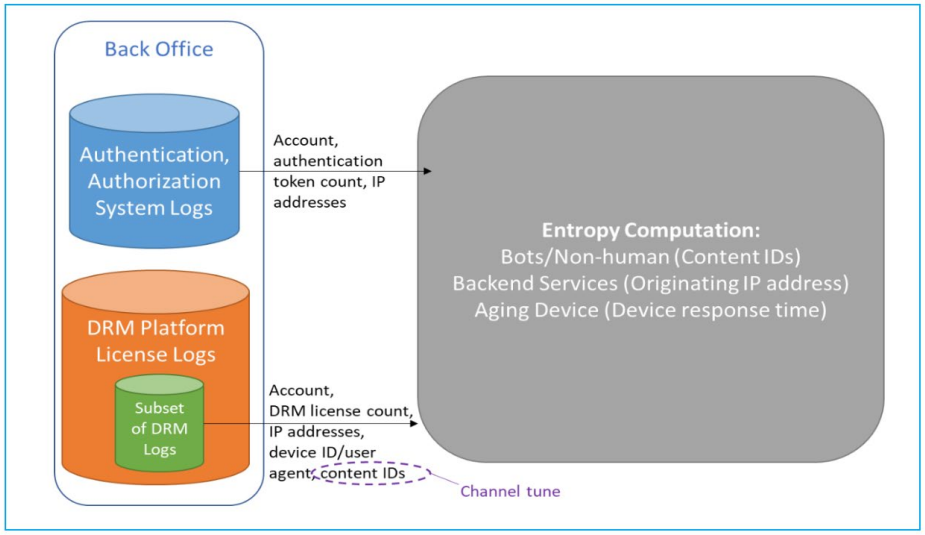

Figure 2 - Entropy Computation Workflow

One of the known issues is when a malicious actor obtains content from service providers illicitly and re-distributes it to unauthorized users. A sign of such automated bot activity is when a large number of DRM license requests are received, exceeding the normal usage levels. While this is an identifier of bot activity, a deeper analysis is warranted as illustrated below.

2.1. Complexity of Bot Analysis

Our studies have shown that bots created for different purposes exhibit different characteristics. These can be analyzed using entropy diagrams. In the following example, the behaviors of three types of bots are compared using an entropy vs. channel count plot. As shown in the diagram below, each bot occupies a different region based on their behavior. The stealth bot is perhaps the most pernicious as it is hard to gauge its extent of damage.

1) **Phishing Bot – objective is to programmatically obtain all the DRMs in a sequential manner**
   CC Sequence: 1,2,3, … ,98,99,100 - CC total count=100
   Shannon Entropy=6.64

2) **Proxy Bot – objective is to obtain specific DRMs on behalf of their users' demand**
   CC Sequence: 1,1,1,1,1,1,1,1, … 1, 2,2, … 2,2 (eighty 1s and twenty 2s) - CC total count=100
   Entropy=.722

3) **Stealth Bot – objective is to obtain a limited number of targeted DRMs and remain un-detected**
CC Sequence: 1,2 - CC total count=2
Entropy=1.0

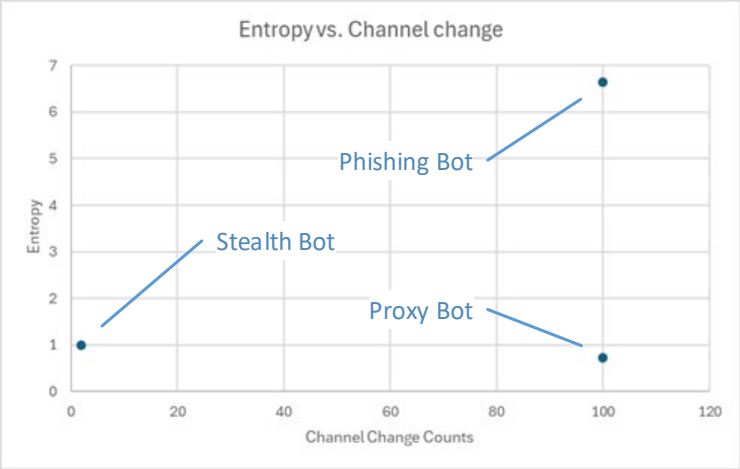

Figure 3 - Behavior of Different Bot Types

2.2. Identifying Network Induced Errors

In addition to bots, the solution presented is capable of proactively identifying anomalies in video delivery components/systems before customers call in to report potential issues. Two examples are cited below.

Backend services – When authentication systems take longer than the average/expected time to respond to customer authentication requests (i.e. high latency), that might indicate the backend services are not scaled properly to address increasing traffic through organic growth and/or peak TV viewing time. When unchecked, this will impact customer experience.

Devices that need attention/aging devices – As new devices are deployed, aging devices are sometimes not churned out of the systems/network. This could be due to customer reluctance to upgrade an aging device and/or unwillingness to learn to use a newer device. In some cases, the device behavior/performance does not line up with the newer devices, even within the same brand.

3. Entropy Algorithms

This section covers the entropy algorithms used in the paper.

3.1. Shannon Entropy

Entropy has its roots in physics and statistical mechanics, where it denotes the disorder or randomness of a physical system. Claude Shannon introduced the entropy concept in his formulation of Information Theory (1948) to quantify the amount of information in a set of random outcomes.

Given the probabilities 'P' of a random distribution 'X', the informational entropy 'H' is given by,

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

The summation is carried out over all possible outcomes. If the outcome of an event is more likely, the entropy value 'H' will be low. On the other hand, if the dataset is more disordered, then the outcome will be hard to predict (more uncertainty). In such a scenario the calculated entropy will be high.

As there are millions of license requests per day, it is not practical to perform this evaluation in an automated fashion. Gini index (below) offers a fast validation mechanism that can be automated in a large network.

3.2. Gini Impurity/Index

Named after the Italian statistician Corrado Gini, it is a measure of purity of elements in a class in machine learning (decision trees). If all elements belong to one class ('pure' scenario), then the Gini index is 0. It reaches the highest value of 1 when the mix is completely random.

$$Gini(E) = 1 - \sum_{j=1}^{c} p_j^2$$

Gini index can be seen as the probability of sampling two observations of different classes in a dataset. For example, for a homogeneous data set (no impurities) such probability will be 1 (i.e. 100%). Unlike the Shannon equation, the Gini formula is faster to compute as it does not contain logarithms (see Reference [2]). This is important when developing an automated solution to handle tens of millions of devices.

While the proposed solution has adopted Gini index for its computational efficiency, the methodology disclosed is equally applicable to other entropy-based methods including classical Shannon formula and its many variations (see reference [3]).

Note – Gini impurity/index in data science is different from the widely known Gini coefficient/index in socioeconomics. It is also named after the same inventor but serves a different purpose (see reference [Gini coefficient]).

4. Methodology

Average users exhibit consistency in their channel tuning behavior. For example, if the user mainly watches news and sports channels, the channel change sequence could be FOX, CNN, FOX, ESPN, FOX, CNN, FOX, CNN, FOX, ESPN. Each such sequence has an associated entropy/impurity value which we define as **channel diversity** (CD). It is calculated using the Gini impurity/index formula (see Entropy Algorithms section).

Example:

Channel change sequence: FOX, CNN, FOX, ESPN, FOX, CNN, FOX, CNN, FOX, ESPN.

This sequence has 10 events and ESPN occurs twice; hence its probability, (*p-ESPN*) is 2/10.

From the formula,

Gini Index = $(1 - p\text{-}ESPN^2 - p\text{-}CNN^2 - p\text{-}FOX^2)$

$= (1 - (2/10)^2 - (3/10)^2 - (5/10)^2)$

$= 0.62$

We posit that the Gini index is a measure of consistency in channel change behavior. That consistency, however, breaks down under anomalous conditions (such as shared/stolen passwords). In such a situation, many more random channels will be present in the sequence. As the channel sequence becomes more **diverse** this change is reflected as a higher Gini index.

In general, abnormally high diverse channel tunes can be attributed to several factors:

1) Automated bots

2) Outliers (channel zapping - unhappy customers that simply surf the channels)

3) Network errors

The channel diversity as defined above is an inherent marker of viewing behavior and therefore an indication of the number of users behind a unique entry-point. We define the latter attribute as **multiplicity**, which could range from a few individuals to hundreds or thousands of virtual entities as in the case of an automated bot. Note that the multiplicity could be a qualitative or quantitative measure. Also, the 'users' in this context could be real/virtual. Examples of 'entry-point' to the network are the user account or device IP/MAC address.

As an example, a single user or a single device household could have a channel change sequence of A-B-C-B-A. However, a sequence with contiguously repeated channels (e.g., A-B-B-C-A) indicates an anomaly (more than one user or network error).

4.1. Metrics

DRM governs the legal access to digital content. When a user device tunes to a channel, it obtains the decrypt key from the license server (via a sophisticated mechanism which is not relevant). The sequence of such DRM license requests/grants has an associated entropy which is reflected in the Gini index. If the DRM process is compromised, however, then the Gini indices for license requests and grants would differ. We recommend the following Gini indices:

- Gini–DRM-token
- Gini–DRM-license-request
- Gini–DRM- license-grant

For example, if 'Gini–DRM-token' value is different from 'Gini–DRM-license-request,' that might indicate a network error, such as incorrect logging mechanisms or server misconfigurations. It could also be due to bots stealing the credentials, but those instances can be identified and ruled out in further analysis. Other identifying signs are a large number of channel changes in millisecond (or sub-millisecond) durations, which is quite different from human behavior. Network errors can also be cross validated by comparing data from other network sources.

Note that the above measurements are based on channel change behavior and not on channel view time ('watch time variability'). Due to inaccuracies inherent in the data measurement process, it is not used in our calculations except for occasional comparisons for data integrity.

5. Inferring Multiplicity from DRM Data

Multiplicity is defined as the number of users (humans or automated bot), behind a single entry-point of the network. This is generally an account identifier (ACCT) or an IP address.

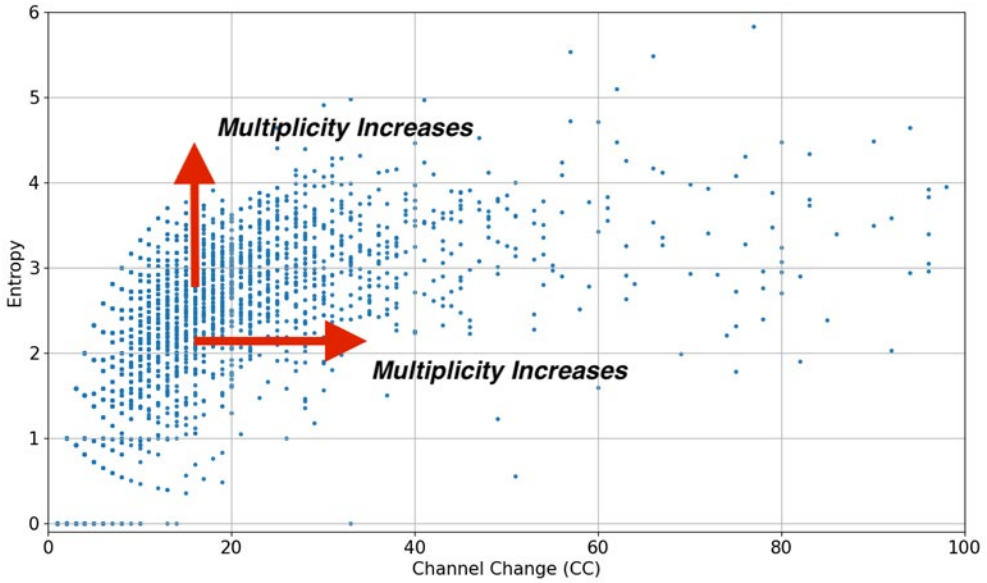

Figure 4 - Multiplicity Directional Variation

Figure 4 is a plot of the channel diversity vs. the channel counts. The former is derived from the Gini formula and the latter is the measured/counted channel changes during the observation period. Each blue dot signifies an entry-point IP address, which could be a single person, household, MDU (multi-dwelling unit), or even automated bot. Multiplicity is the predicted likelihood of multiple users behind a single entry-point to the network. It increases as we traverse rightward along the horizontal axis (more channel counts). The same behavior is observed in the northward direction as well (high channel diversity).

To methodically examine each case, we divide the graph into many regions as shown below in Figure 5.

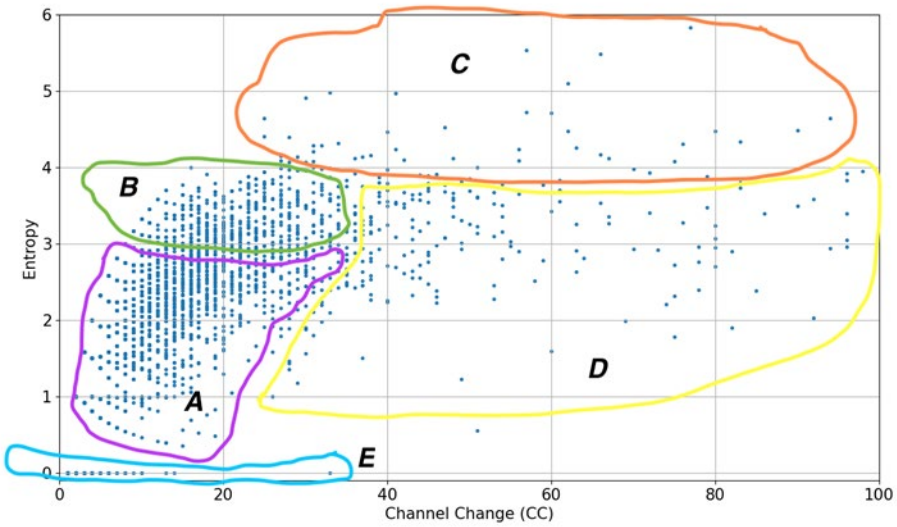

Figure 5 - Region Analysis

Region A – Normal behavior. Low CD values denote a single user or a household. High CD values indicate multiple users sharing the credentials (which could be content/device ID or IP).

Region B – Normal behavior, but high propensity for shared passwords instances.

Region C – This region indicates high number of users such as large MDUs. It might also indicate instances of 'channel zapping' (unhappy customers) or bots.

Region D – Bots (low entropy with repeated Content_ID). It could also represent smaller MDUs.

Region E – Single daily channel selections with no diversion. E.g., sports bar clients and bots dedicated to a single channel.

In the above 'snapshot view,' bots and MDUs (e.g., 50- 100 users in one account) both would exhibit similar behavior. To distinguish between the two, residual analysis is recommended.

5.1. Residual Analysis

We compare two consecutive plots computed at two similar points in the timeline, (e.g., daily at 7PM or two consecutive Sundays). The assumption is that under steady-state conditions, the blue dots should return close to their original positions at consecutive times. By subtracting the coordinates of datapoints from consecutive plots, we obtain a 'residual plot.' Since MDUs are expected to be in steady state, there should not be much movement. In contrast, bots would be active with frequent movement. This will be reflected in the residual plot as long vectors.

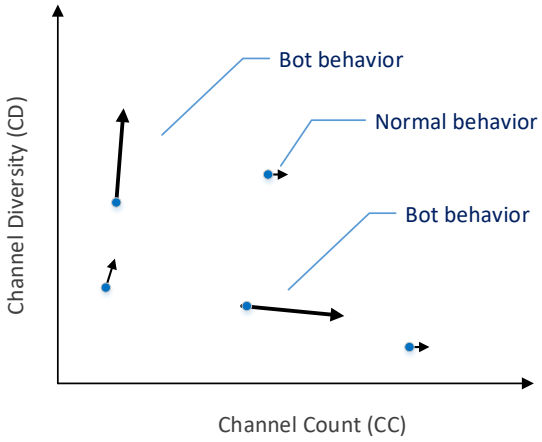

Figure 6 - Residual Analysis

In Figure 6, the location, length and direction of the 'residual vectors' indicate the type of bot and its impact. For instance, a more vertically inclined vector could be an indication of stolen credentials for a major sports event (no increase in channel count). Similarly, a horizontally inclined vector might be indication of someone selling/sharing credentials indiscriminately at scale.

The above vector analysis could provide further insight into time-series evolution. Mathematically, the affinity of two vectors is compared with their cosine similarity and 'scalar product' measures. By studying the long behavior of above vectors, different types of bots (and other miscreants) can be identified based on their similar characteristics.

5.2. 3D Plots and Layered View

In addition to channel count (x-axis) and channel diversity (y-axis), a third dimension could be added to gain further insights on behavioral patterns. As an example, content of each channel viewed belongs to a TV **genre**. There are over a dozen such categories (news, sports, talk shows, game shows, sitcoms, reality, drama, soap, cartoons, etc.). A 3D plot with genre as the z-axis would indicate the prevalence of blue dots by genre as shown in Figure 7. This data would be useful for identifying advertising opportunities, for instance.

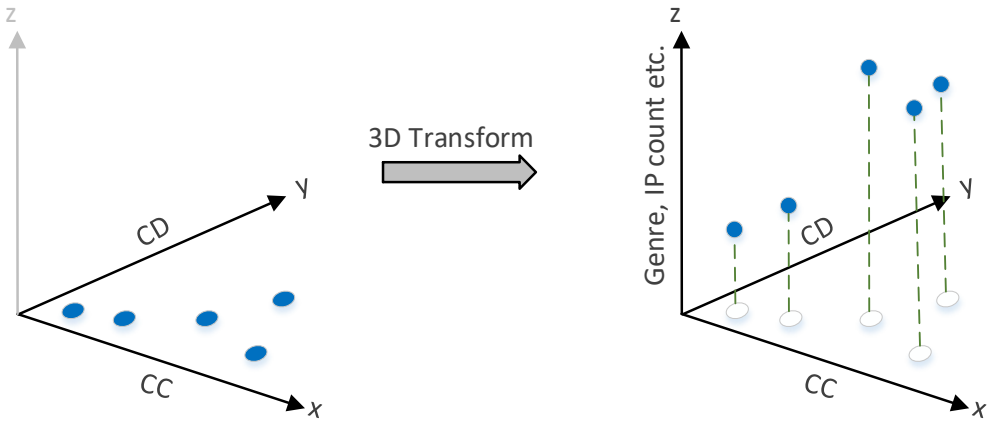

Figure 7 - Layered View for Visualizing Relationships

The diagram below depicts how the 3D transform would function with real data.

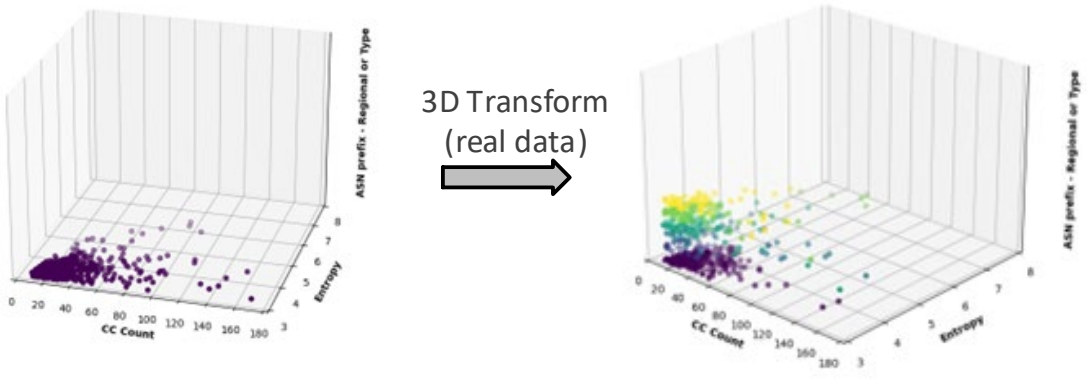

Figure 8 - 3D Transform Depicting Multiple ASN Subnets (IPs) in Entropy vs. CC Plots

Other choices for the z-axis are how long one stayed as a customer, or any other demographic feature, such as income or age. A machine learning classification study can be performed with historical data to understand the relation among the variables (e.g., which demographic is more correlated with channel-zapping).

Plotting the IP address (or ASN) data for the z-axis enables us to visualize relationships useful in troubleshooting (e.g., which origination points are more susceptible to network errors).

5.3. Targeted Advertising Opportunities

As part of the Gini index calculation, the channel change probability counts are computed for each channel. The channels with high probabilities can be thought of as the most viewed. Conversely, channels with low probabilities are the least viewed. This data would be helpful to identify targeted ad opportunities.

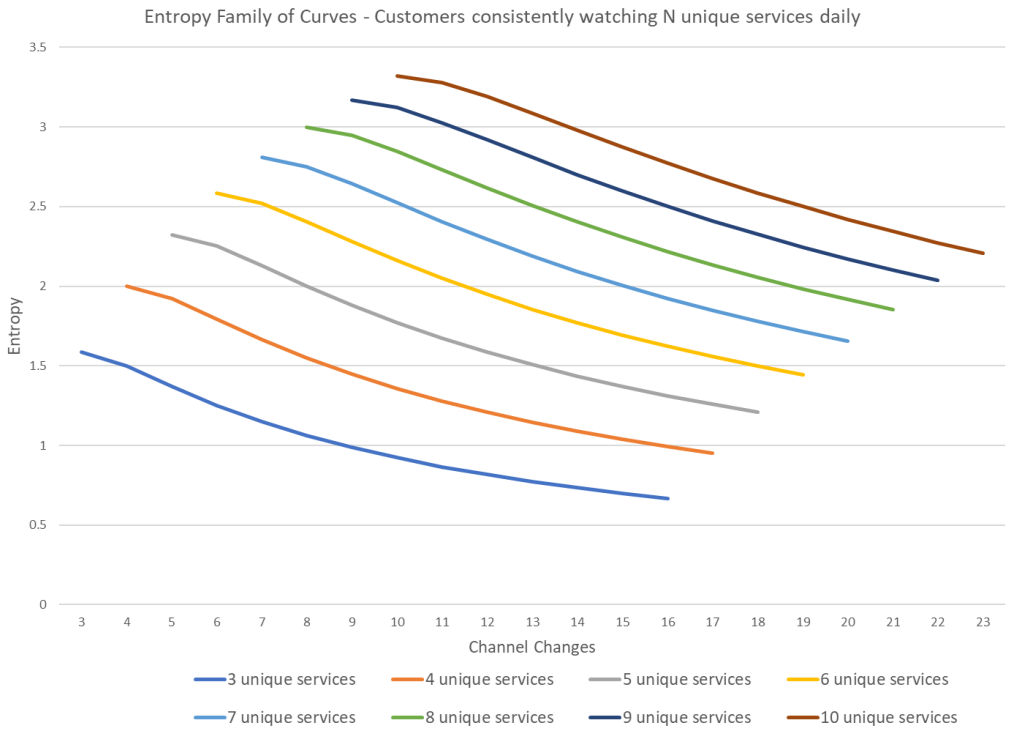

Figure 9 - Identifying the Likelihood of a Customer Being on a Certain Channel

Referring to Figure 9, as the channel change sequence extends out along each sequence curve, it also provides insight into the likelihood of that customer being on a specific channel and watching a targeted ad. For instance, the lowest entropy curve (the blue line or '3 unique services') is a customer device IP consistently watching only three unique services. They are identifiable and have a high likelihood of viewing ads on those three services. Customers with channel change behavior on the bottom curve have a higher probability of seeing a targeted ad on one of their services than customers on the higher entropy curves.

This information is useful when planning a successful ad campaign as the marketer can target the optimal customer base. High entropy would indicate the customer is more likely to be watching a particular channel and the accompanying TV ad. Conversely, low probabilities would indicate low viewership and little value for ad sales.

5.4. Identifying Localized Impacts

The anomalous data points in the 'snapshot view' could be due to bots or network errors. Using 3D plots, it is possible to differentiate between each, as shown in Figure 10. In this case, the demographic data (e.g., DMA or a collection of zip codes), form the z-axis. The premise is that network errors would be regional or multi-region, whereas the bot impact would be distributed indiscriminately. In a 3D plot as shown, the points due to network errors will be confined to different layers, whereas bot behavior would be indiscriminate.

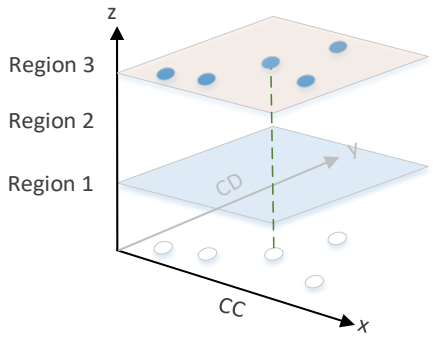

Figure 10 - Layered View Displaying Network Errors.

6. Conclusion

Current fraud analytics are generally based on aggregated trends and are not sufficiently granular. Also, automating entropy-based methods to track tens of millions of devices are computationally prohibitive. The paper presented a novel approach to address these issues.

Abbreviations

Acronym	Definition	Notes
ACCT	account	
ASN	autonomous system number	IP prefix for a group of IPs
CC	channel change	
CD	channel diversity	
CV	channel view time	
DRM	digital rights management	
DMA	designated market area	
MDU	multi dwelling unit	
NOC	network operations center	

Bibliography & References

[1] "*Gini coefficient*" – https://en.wikipedia.org/wiki/Gini_coefficient

[2] "*Theoretical comparison between the Gini Index and Information Gain criteria", Raileanu et al.* Annals of Mathematics and Artificial Intelligence 41: 77–93, 2004. (A comparison of Gini and Shannon algorithms) – https://www.unine.ch/files/live/sites/imi/files/shared/documents/papers/Gini_index_fulltext.pdf

[3] "*An Entropy-Based Approach for Anomaly Detection",* Aadel Howedi *et al*., MDPI, Published: 30 July 2020. (Different variations of entropy formulae) – https://pubmed.ncbi.nlm.nih.gov/33286616/