Karlsruhe Institute of Technology

KASTEL

Could Network Information Facilitate Address Clustering in Bitcoin?

Till Neudecker, Hannes Hartenstein

Institute of Telematics,
DSN Research Group, Prof. Hartenstein

Motivation

“The first node to inform you of a transaction is the source of it”

Dan Kaminsky, BlackHat 2011 - Black OPS of TCP/IP

Motivation

“The first node to inform you of a transaction is the source of it”
Dan Kaminsky, BlackHat 2011 - Black OPS of TCP/IP

Dan Kaminsky’s disclaimer: “more or less true, and absolutely over time”

Reasons why this might <u>not</u> work:

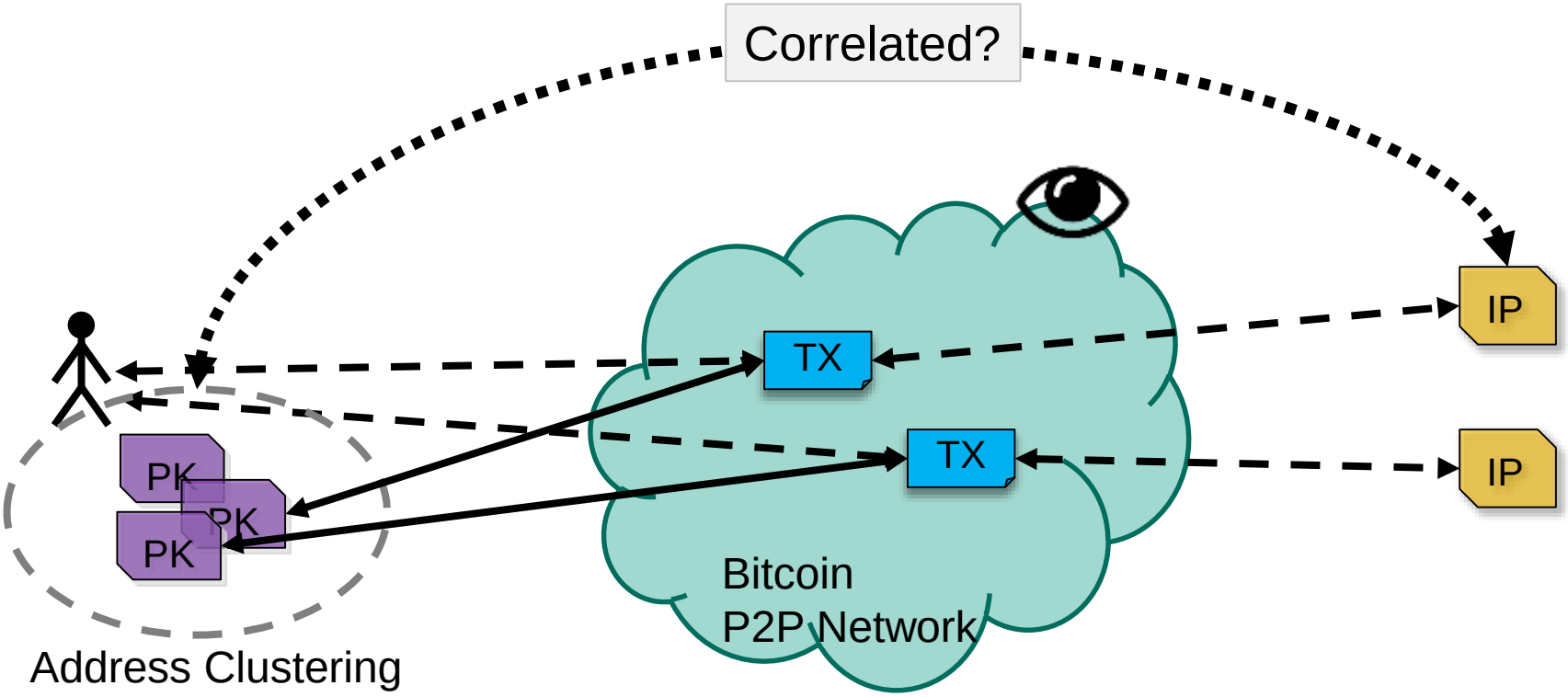
- Peers behind NATs
- Dynamic IP addresses
- Trickling of transactions
- Reachable peers not operated by “users”
- Users are roaming
- …

Reasons why this might work:

- Deanonymisation of clients in Bitcoin P2P network. Biryukov et al. (CCS‘14)
- An Analysis of Anonymity in Bitcoin Using P2P Network Traffic. Koshy et al. (FC’14)
- Bitcoin over Tor isn't a Good Idea. Biryukov et al. (SP’15)
- Some information still leaks over time …

? Is there a correlation between “the first node to inform you” and the user issuing the transaction ?

Approach


Correlated?
IP
TX
TX
IP
PK
PK
PK
Bitcoin
P2P Network
Address Clustering

Address Clustering

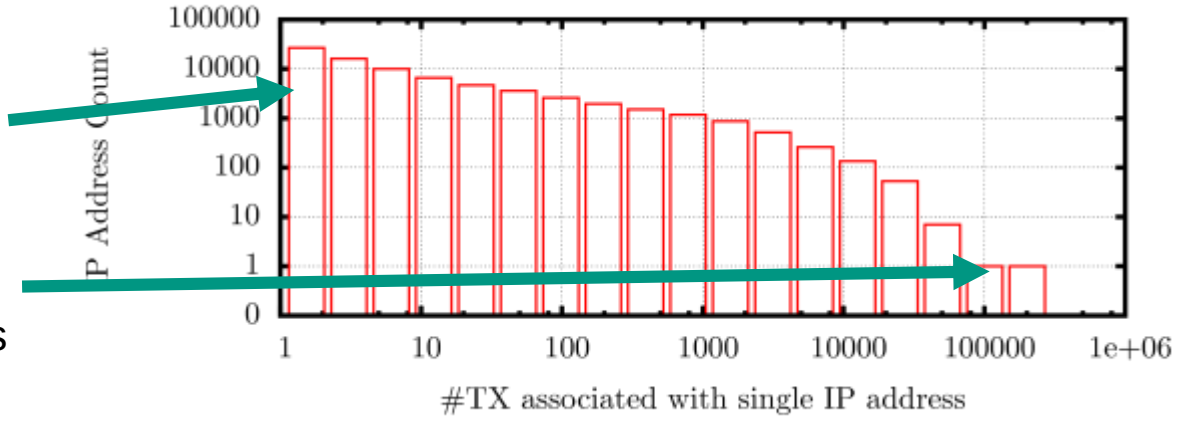
- Many heuristics published (e.g., Reid et al. 2013, Meiklejohn et al. 2013)
  - Multi-Input
  - Several change address variants
  - Value based
  - Growth based
  - …
- Which one to use?
  - **We don’t know which heuristics really works best**
  - Simply try all… as long as any one works, we are fine
- In the paper: comparison between them
  - Heuristics differ in many aspects (e.g., size of largest cluster between 0.1m and 85m)
  - Source code available at www.github.com/tillneu/bitcoin-clusterer

TX – IP-Address Association

- Two monitor nodes, running since July 2015, ~ 100m Transactions
  - dsn.tm.kit.edu/bitcoin
  - Initially set up for topology inference of bitcoin P2P network

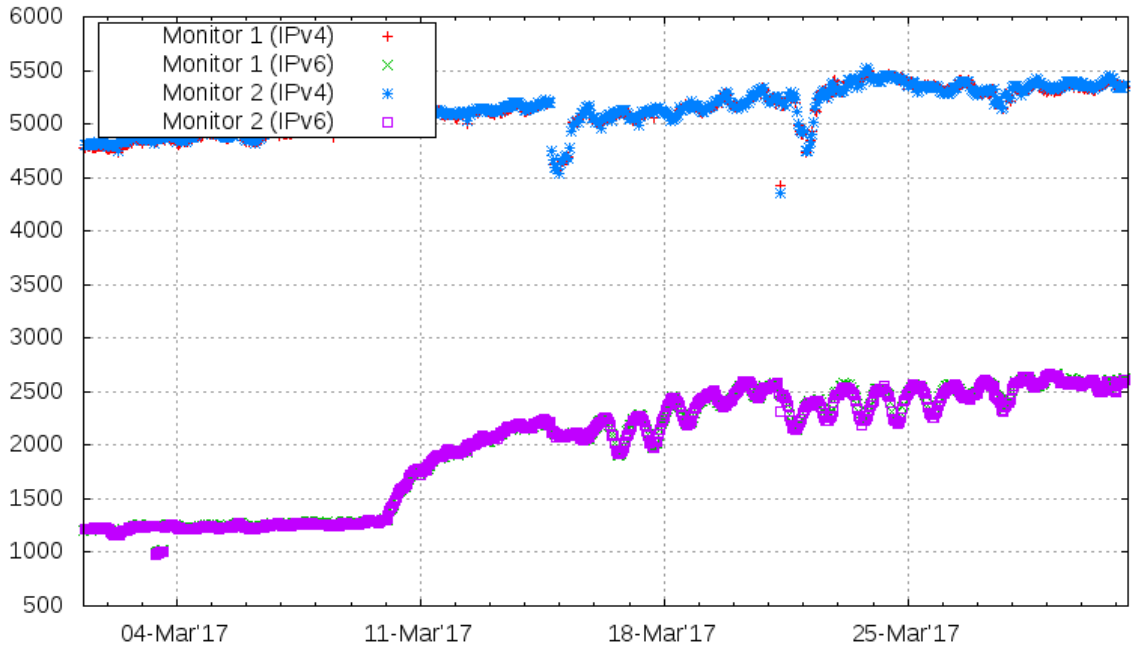
- Strategy: Associate first* IP address to announce TX (via INV) with TX
  - *Subtract estimated latency to foreign peer to get first *sending time*
  - Discard obviously false mappings, e.g.:
    - different first IP address for both monitor nodes
    - subsequent receptions would be faster than the speed of light
  - ~10% of all TXs are associated to an IP address

- Results:
  - Most IP addresses only associated to a small number of TXs
  - Very few IP addresses associated to many TXs



TX – IP-Address Association

- Two monitor nodes, running since July 2015, ~ 100m Transactions
  - dsn.tm.kit.edu/bitcoin
  - Initially set up for topology inference of bitcoin P2P network



Correlation – Contingency Table

	1.1.1.1	2.2.2.2	3.3.3.3
Cluster A	1	2	5
Cluster B	3	2	3
Cluster C	4	1	0

→ ~80k IP addresses

↓ # Cluster: 50m-150m

Cells: How many Transactions were issued by Cluster x and associated to IP address y

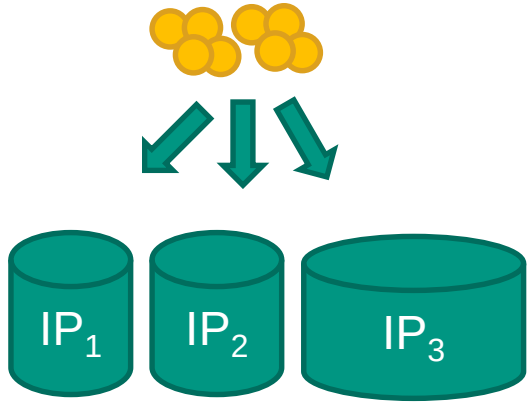
~8 trillion cells but only 10 million TX
→ Too sparse for standard statistics!

Correlation II – Poor Man's Statistics

	1.1.1.1	2.2.2.2	3.3.3.3
Cluster A	1	2	5
Cluster B	3	2	3
Cluster C	4	1	0

$\Sigma = 8$

Corresponding ***balls-into-bins*** experiment:



Size of bin ≙ total number of TX associated with IP address

1. Look at one row (column) at a time
2. Look at the highest value (i.e., the association with the most TX)
3. What is the probability of randomly assigning that many transactions to one IP address, assuming independence of clusters and IP addresses?
4. If that probability is lower than significance level (1%), mark cluster/IP address association as **conspicuous**

Results & Discussion

- Number of Clusters with >= 2 TX w/ IP address: ~282k – 456k*
- Number of conspicuous clusters: ~15k – 36k*

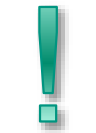
→ Share of conspicuous clusters: ~5% – 8%*

→ Share of conspicuous IP addresses: ~6% – 20%*

*depending on clustering heuristic

- Small number of participants exhibit potentially exploitable behavior
- Causes of these observations unclear

- Correlation only found for small share of clusters / IP addresses
- Only in these cases, network information might be used for clustering



- More powerful adversary might extract more information
- Other statistical methods might reveal more correlations

Table 1. Comparison of all heuristics. Total number of addresses: 196,963,722, total number of transactions: 172,868,721.

Heuristics	# Cluster	∅Size	max size	#clusters w/ size 1
H1	88 m	2.24	12 m	65 m
H1+H2	46 m	4.25	92 m	29 m
H1+H2a	51 m	3.89	87 m	32 m
H1+H2b	63 m	3.10	66 m	40 m
H1+H2c	48 m	4.13	85 m	30 m
H1+HV	72 m	2.71	76 m	62 m
H1+HG$_{10}$	146 m	1.34	0.1 m	123 m
H1+HG$_{100}$	121 m	1.62	0.25 m	97 m
H1+HG$_{1000}$	108 m	1.83	1 m	84 m
H1+HG$_{10000}$	104 m	1.88	8 m	81 m

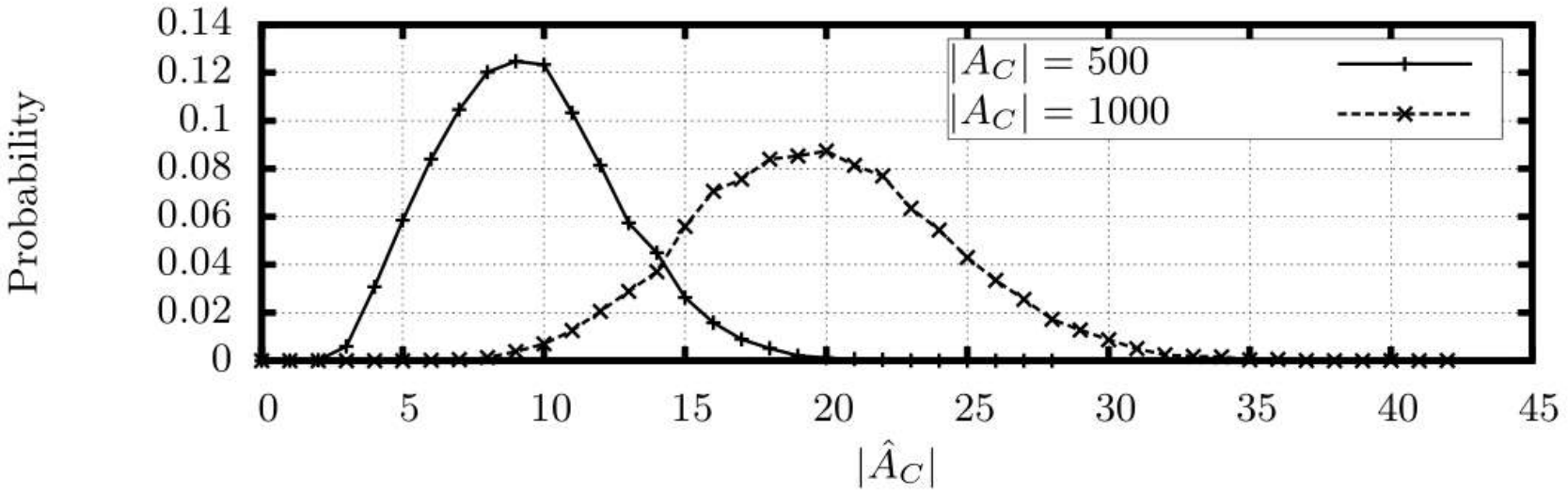


Fig. 3. Probability distribution $P_i(X = |\hat{A}_C|)$ for $|A_C| = 500$ and 1,000 transactions, respectively, assuming independence and given the empirical IP address counts (cf. Fig. 2). Values numerically approximated.

Table 2. Comparison of the number of clusters with at least two associated IP addresses ($|\{C : |A_C| \geq 2\}|$) and the number and share of conspicuous clusters (C^+), and the share of conspicuous IP addresses ($\mathcal{A}^+$) for various heuristics.

Heuristics	$\lvert\{C : \lvert A_C\rvert \geq 2\}\rvert$	$\lvert C^+\rvert$	$\frac{\lvert C^+\rvert}{\lvert\{C:\lvert A_C\rvert\geq 2\}\rvert}$	$\frac{\lvert \mathcal{A}^+\rvert}{\lvert\{A:\lvert T_A\rvert\geq 2\}\rvert}$
H1	282,950	14,879	5.26 %	18.7 %
H1+H2	398,802	32,623	8.18 %	6.2 %
H1+H2a	387,696	32,026	8.26 %	6.2 %
H1+H2b	456,063	35,138	7.70 %	6.5 %
H1+H2c	452,189	35,602	7.87 %	6.7 %
H1+HV	296,132	14,736	4.97 %	6.9 %
H1+HG$_{10}$	299,140	15,537	5.19 %	16.7 %
H1+HG$_{100}$	300,927	15,755	5.23 %	19.6 %
H1+HG$_{1000}$	301,775	16,434	5.45 %	20.2 %
H1+HG$_{10000}$	308,900	18,788	6.08 %	19.7 %

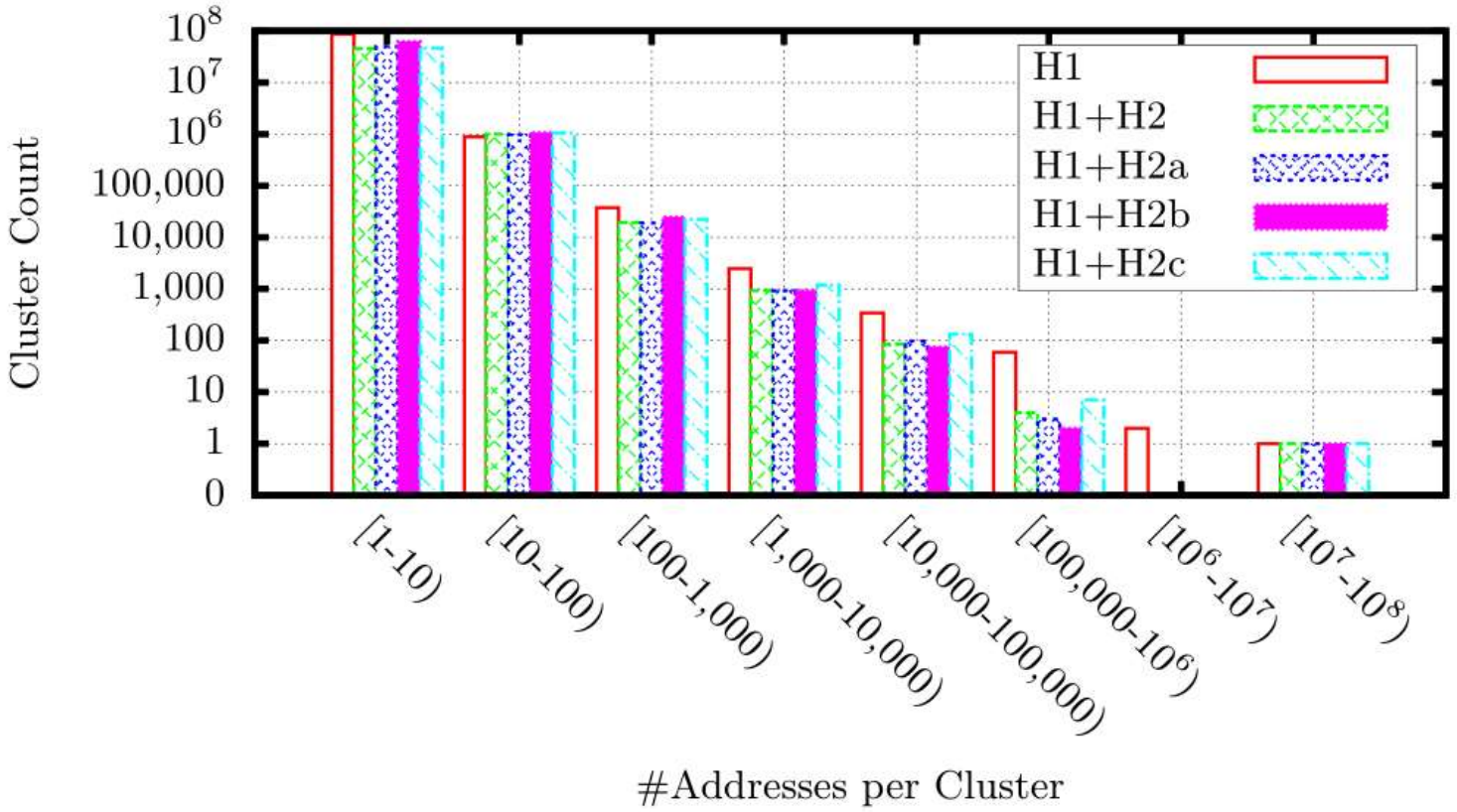


Fig. 4. Histogram of the number of clusters for various sizes (i.e., number of addresses per cluster).

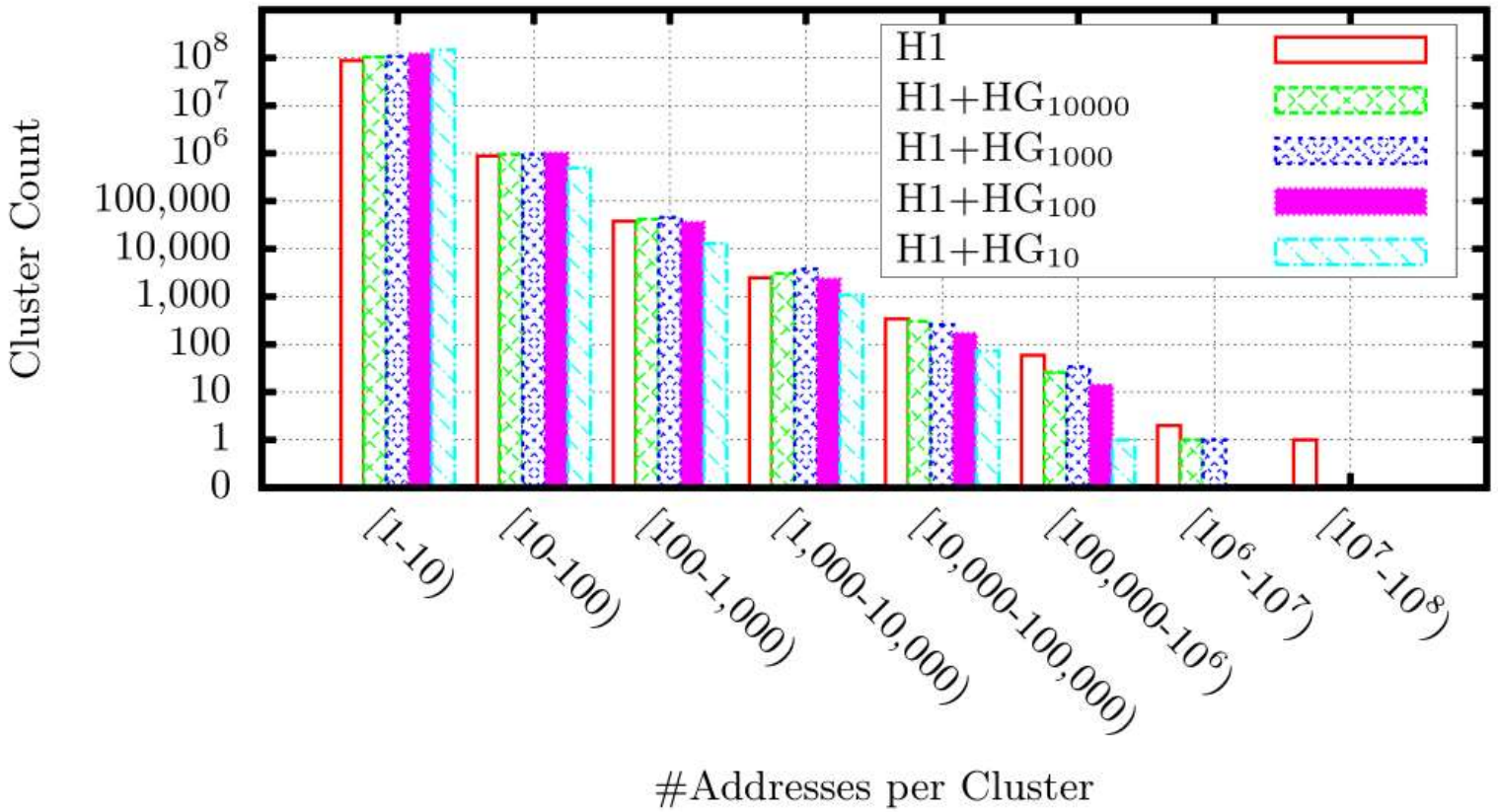


Fig. 5. Histogram of the number of clusters for various sizes (i.e., number of addresses per cluster).