Understanding Block-level Address Usage in the Visible Internet

Xue Cai and John Heidemann

USC/Information Sciences Institute

Aug. 31, 2010, SIGCOMM'10

The Discovery of Halley's Comet

The Discovery of Halley's Comet

2 *historical records*
(year 1531, 1607)
1 *observation*
(year 1682)

Edmond Halley

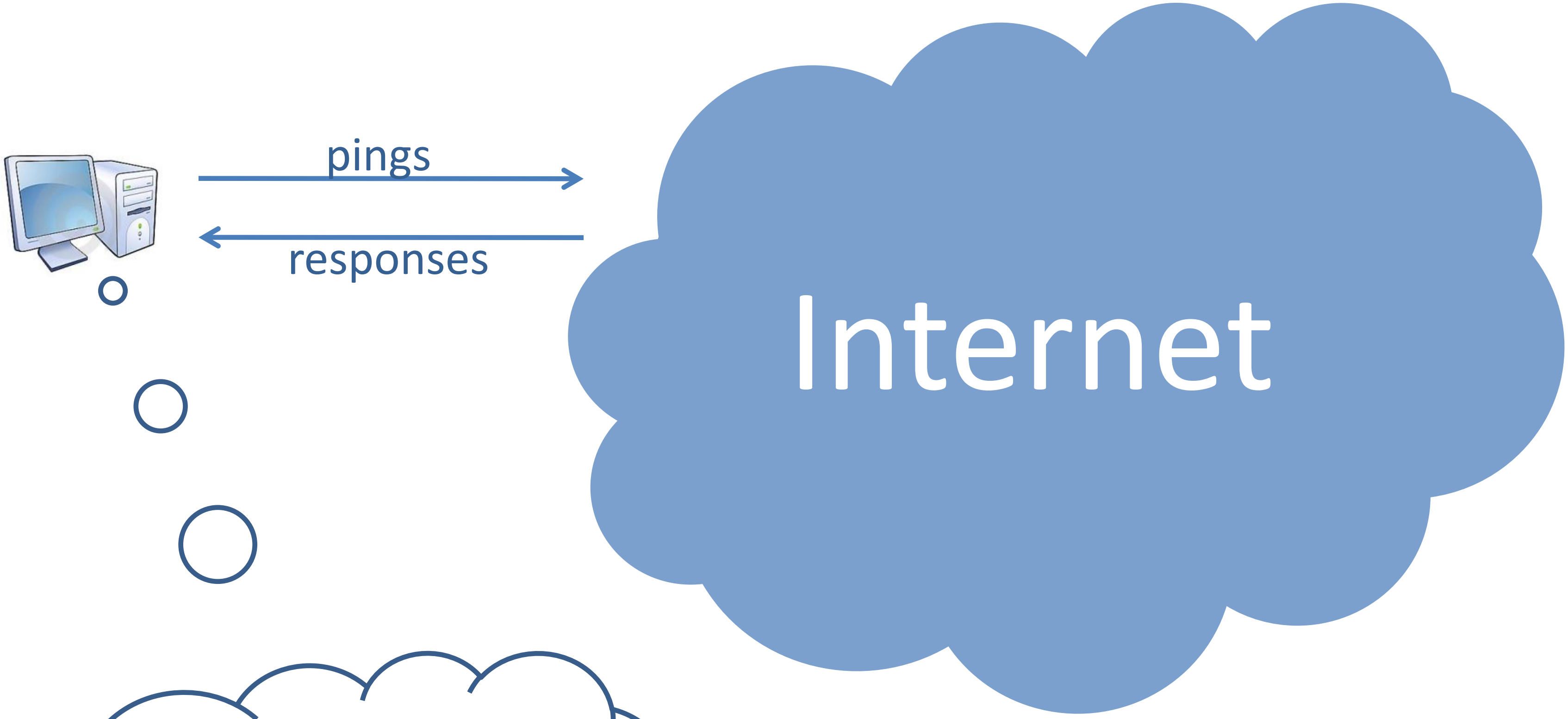
"It's the same object which returns to earth every 76 years. "

3 simple observations

an astronomer

1 simple characteristic of the comet

SIMPLE observations inferred **SIMPLE** conclusion can have **TREMENDOUS** value.

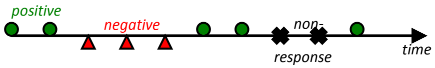


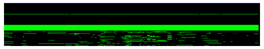
Address Utilization?
Dynamic Addressing?
……

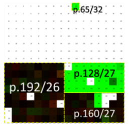
Our Q: what can simple observations about the Internet say?

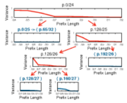
Key Contributions

Methodology



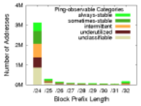


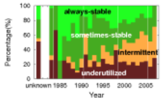




- Active probing, pattern analysis, clustering, classification

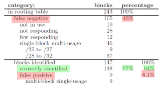
Application



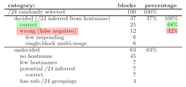


- Network management, resource allocation, Internet trend study

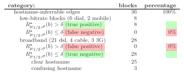
Validation



category:	blocks	percentage	
in routing table	243	100%	
false negative	105	43%	
not in use	19		
not responding	28		
few responding	12		
single-block multi-usage	46		
/25 to /27	9		
/28 to /32	37		
blocks identified	147		100%
correctly identified	138	57%	94%
false positive	9		6.1%
multi-block single-usage	9		



category:	blocks	percentage	
/24 randomly selected	100	100%	
decided (/24 inferred from hostname)	37	37%	100%
correct	25		68%
wrong (false negative)	12		32%
few responding	6		
single-block multi-usage	6		
undecided	63	63%	
no hostname	45		
few hostnames	7		
potential /24 inferred	7		
correct	7		
has sub-/24 groupings	4		



category:	blocks	percentage
hostname-inferrable edges	36	100%
low-bitrate blocks (6 dial, 2 mobile)	8	
$R^*_{\mu_{1/2},\sigma}(b) > \delta$ (true positive)	8	
$R^*_{\mu_{1/2},\sigma}(b) \leq \delta$ (false negative)	0	0%
broadband (21 dsl, 4 cable, 3 3G)	28	
$R^*_{\mu_{1/2},\sigma}(b) > \delta$ (false positive)	0	0%
$R^*_{\mu_{1/2},\sigma}(b) \leq \delta$ (true negative)	28	
clear hostname	25	
confusing hostname	3	

- USC’s network, the general Internet, consistency across time

Key Contributions

	Methodology	Application	Validation
Spatial Correlation?	Group addresses into blocks by usage	More frequent probing? Block sizes? Block-level usage?	USC's network, General Internet, Consistency
Address Utilization?	Find blocks with less than 10% time responsive	Resource reallocation? Efficient management?	USC's network, General Internet, Consistency
Dynamic Addressing?	Blocks switching state (up/down) frequently	Botnet detection? Spam filtering? Click fraud?	USC's network, General Internet, Consistency
Low-bitrate Identification?	Utilize standard deviation of RTTs	Auto content serving? Network management?	USC's network, General Internet

Key Contributions

	Methodology	Application	Validation
Spatial Correlation?	Group addresses into blocks by usage	More frequent probing? Block sizes? Block-level usage?	USC's network, General Internet, Consistency
Address Utilization?	See paper	See paper	See paper
Dynamic Addressing?			
Low-bitrate Identification?	Utilize standard deviation of RTTs	Auto content serving? Network management?	USC's network, General Internet

Related Work

- J. Heidemann, Y. Pradkin, R. Govindan, C. Papadopoulos, G. Bartlett, and J. Bannister. Census and Survey of the Visible Internet. In *Proceedings of the ACM Internet Measurement Conference (IMC)*, p. 169-182. Vouliagmeni, Greece, October, 2008.

- What's the same?
  - Collection methodology (and datasets)
  - Error bounds on ping census accuracy: undercounts by about 40%
  - Preliminary metrics

- What's new? *deeper understanding; new interpretation*
  - ***new metrics***
    - block-level analysis, not just addresses
    - RTT, not just responsivness
  - ***new algorithms***
    - block identification
    - low-bitrate identification
  - ***new conclusions***
    - evaluation of block utilization
    - trends of address utilization
    - trends of dynamic addressing

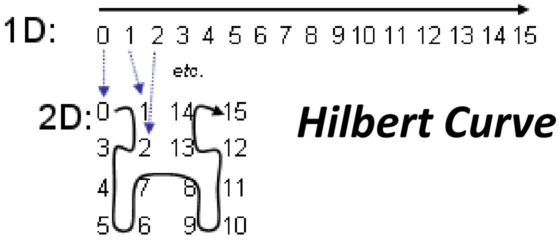
Key Contributions

	Methodology	Application	Validation
Spatial Correlation?	Group addresses into blocks by usage	More frequent probing? Block sizes? Block-level usage?	USC's network, General Internet, Consistency
Address Utilization?	**See paper**	**See paper**	**See paper**
Dynamic Addressing?			
Low-bitrate Identification?	Utilize standard deviation of RTTs	Auto content serving? Network management?	USC's network, General Internet

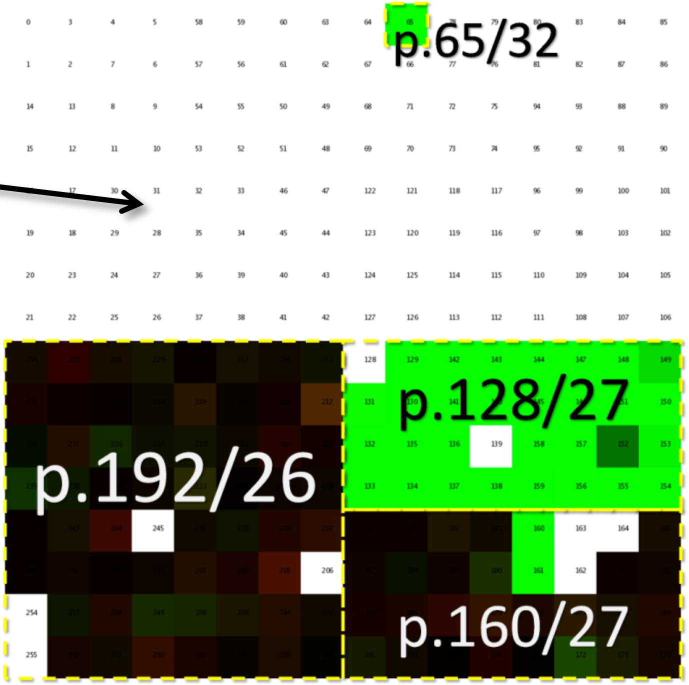
Background: What space?

- IPv4 address space
  - *address block*: *p/n:* addresses with common *n*-bit prefix *p*
  - *a.b.c.d* and *a.b.c.*(*d*+1) are *adjacent addresses*

A /24 block (*p/24*) with 256 addresses,
Layout ***Hilbert Curve*** keeps **adjacent addresses physically near each other.**



Hilbert Curve



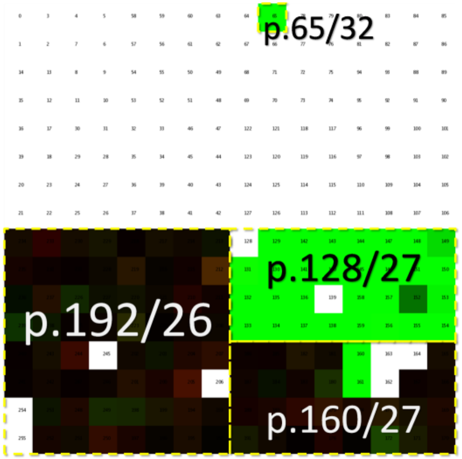
Hypothesis: Spatial Correlation

- What is Spatial Correlation?
  - adjacent addresses are likely to be used in the same way

  ⇒ **spatial correlation of address blocks**

  ⇒ **usage blocks**

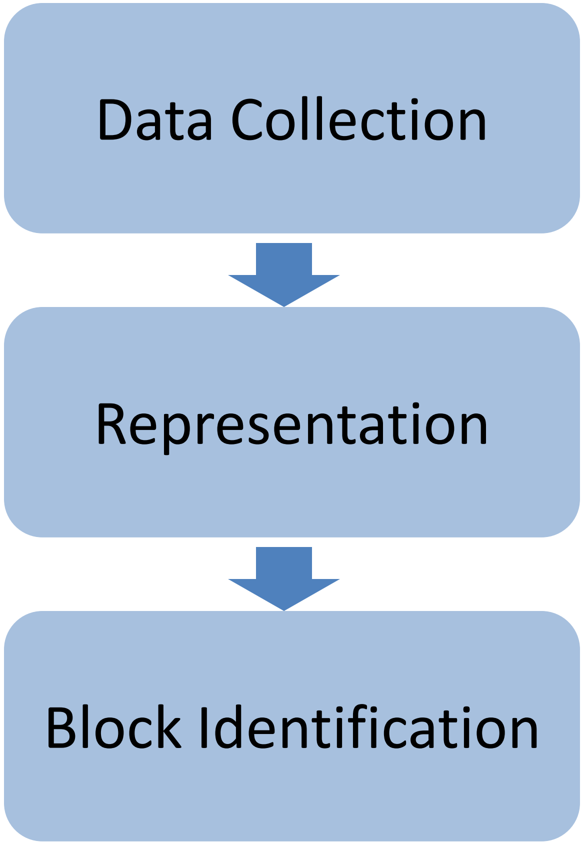
- Usage blocks
  - are NOT **allocated blocks**, but correlated
    - Internet addresses are allocated in blocks (ICANN to regional registries to ISPs to you)
    - addresses in one block are usually assigned to similar users
  - are what we want to observe if exist
    - **observable blocks** → usage blocks



Spatial Correlation: Application

- Why care?
  - Efficiently select **representative addresses** to conduct more detailed study
    - Addresses in one block are used in the same way
    - So only need few representatives to probe in the future

Spatial Correlation: Methodology



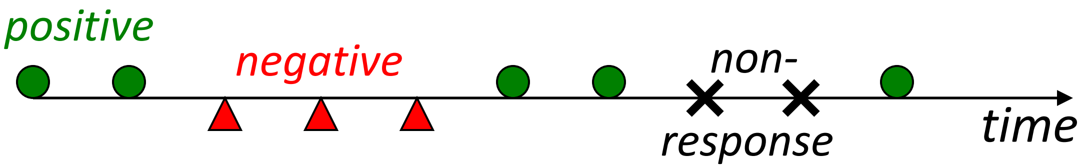
Input: data for individual ***addresses***

Output: address sharing similar usage grouped into ***observable blocks***

Spatial Correlation: Data Collection

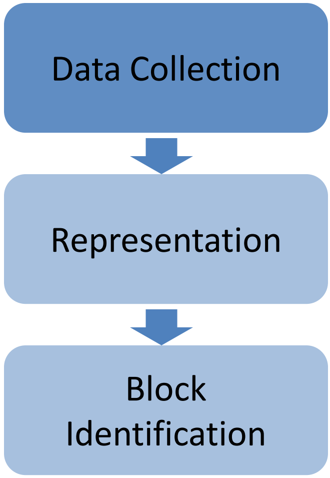
How **Ping** each address in random /24 blocks every 11 minutes for a week and **collect the probe responses**.

1% of the allocated IPv4 address space probed.

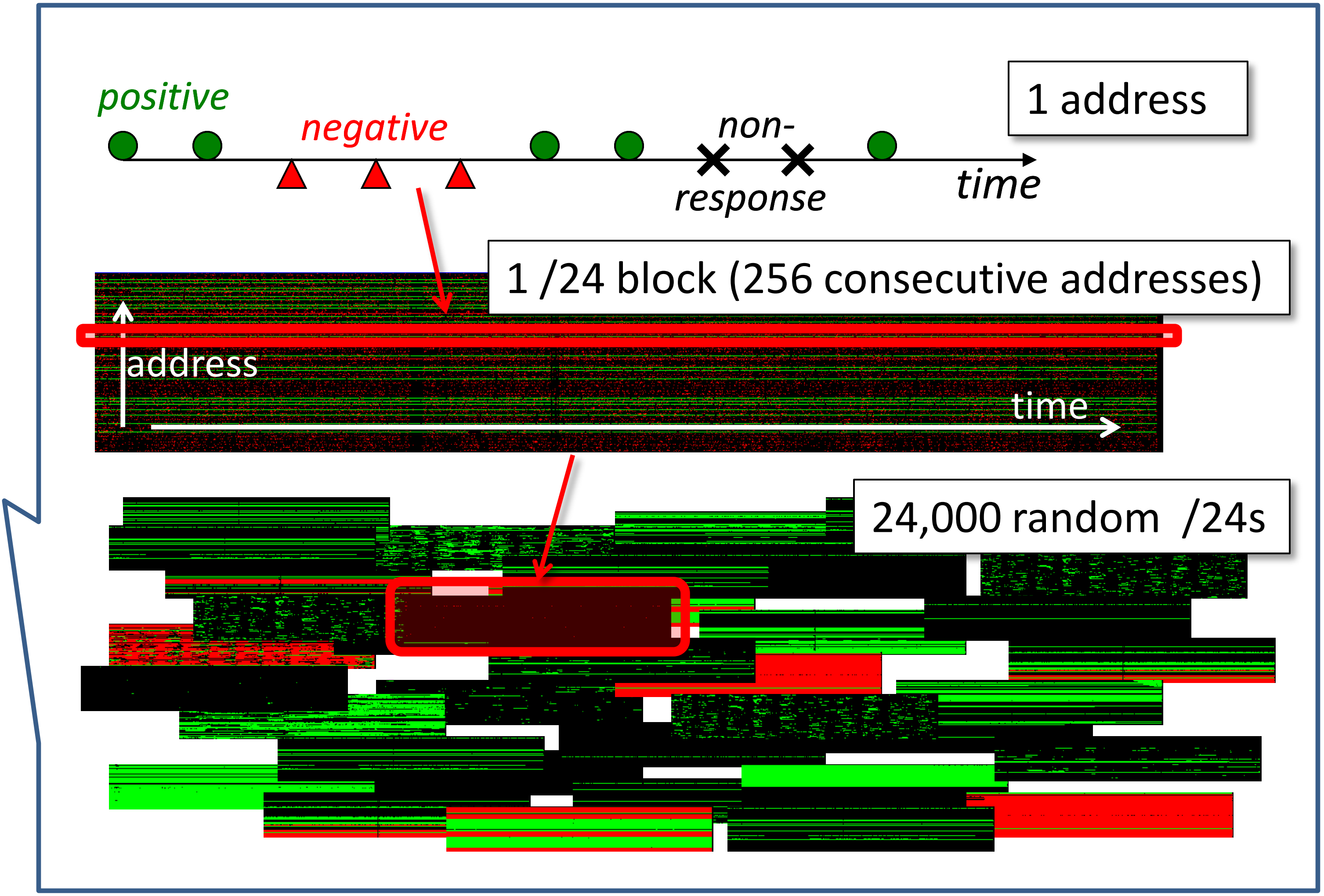


Why Systematic pings reveal more information.

Validity of ping: IMC’08 paper established error bounds: not perfect, but often pretty good; ~40% undercount



Spatial Correlation: Data Collection



Data Collection

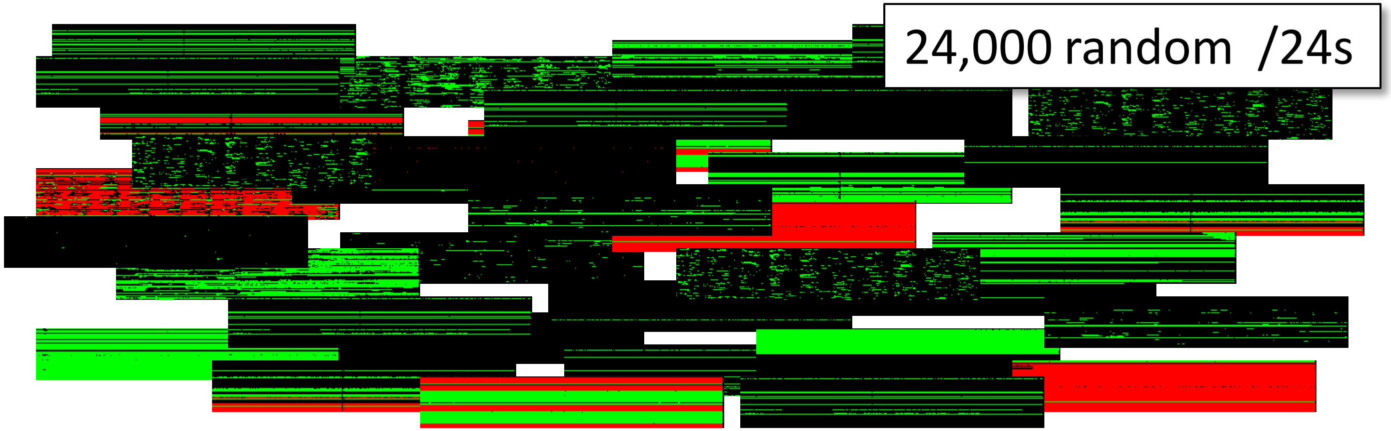
Representation

Block Identification

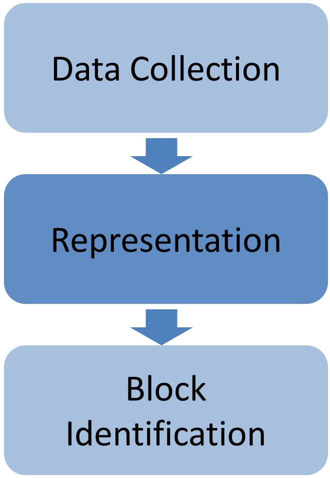
Spatial Correlation: Representation

Why

One survey: > 5 billion ping responses, **need more meaningful representation to represent address usage**



24,000 random /24s

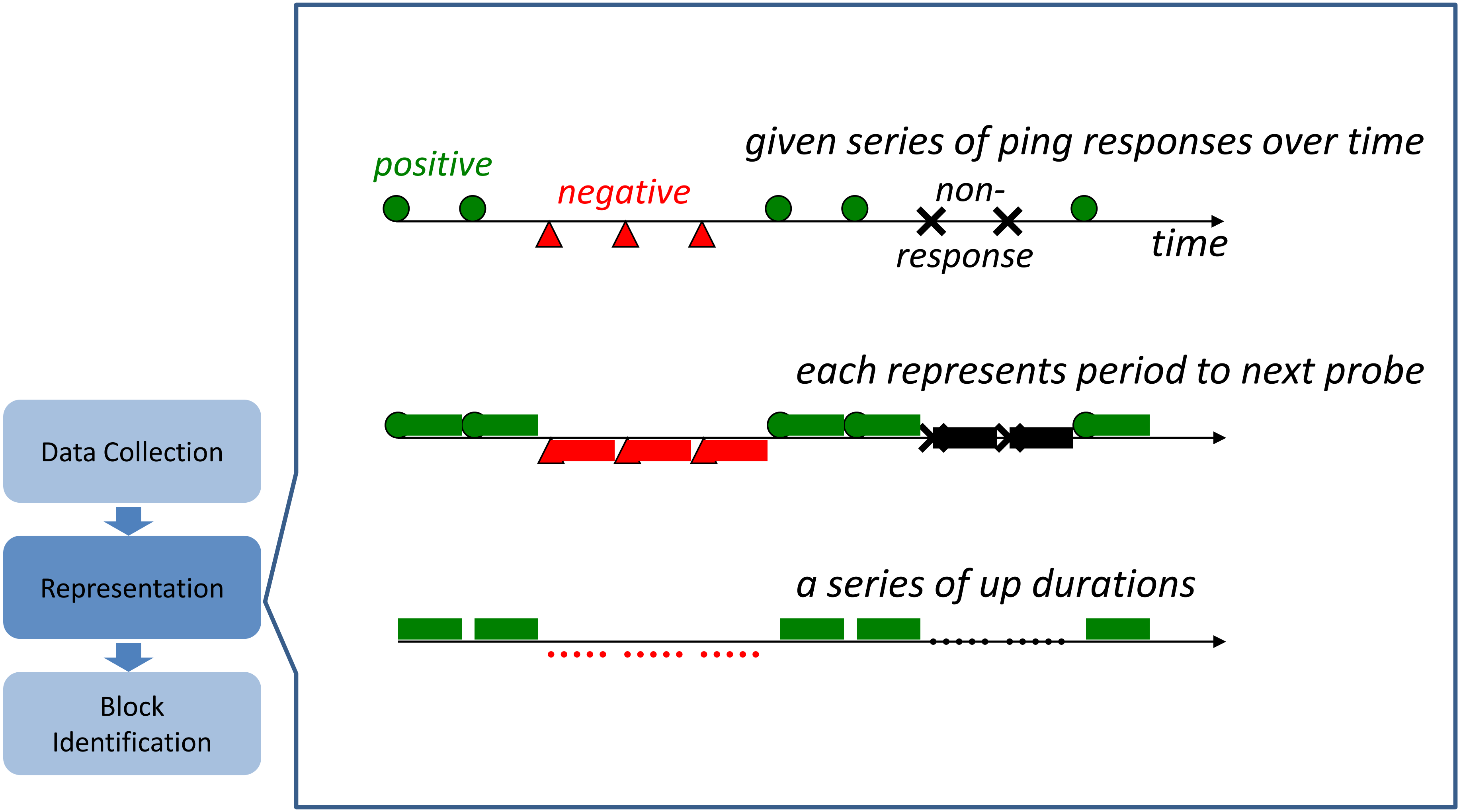


Data Collection

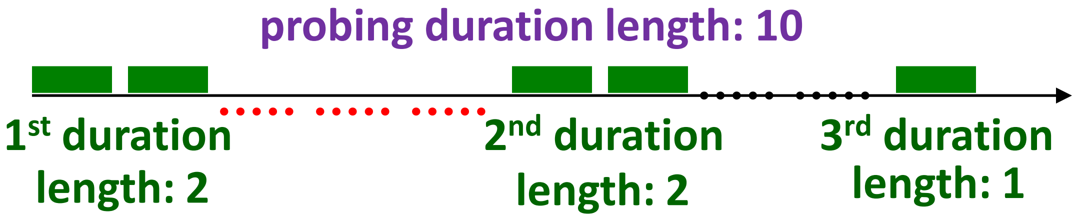
Representation

Block Identification

Spatial Correlation: Representation



Spatial Correlation: Representation



How *3 metrics to capture address usage*

Availability (A)
:= normalized *sum* of up durations

Example:
= (2+2+1) / 10 = 0.5

Intuition:
utilization efficiency

Volatility (V)
:= normalized # of up durations

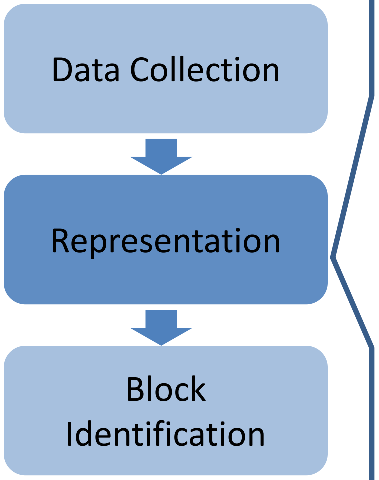
Example:
= 3 / (10/2) = 0.6

Intuition:
high V infers dynamics

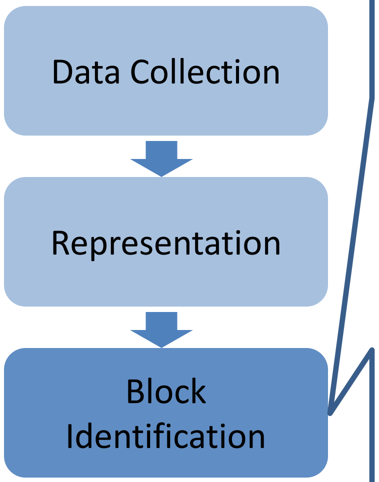
Median-Up (U)
:= *median* up duration

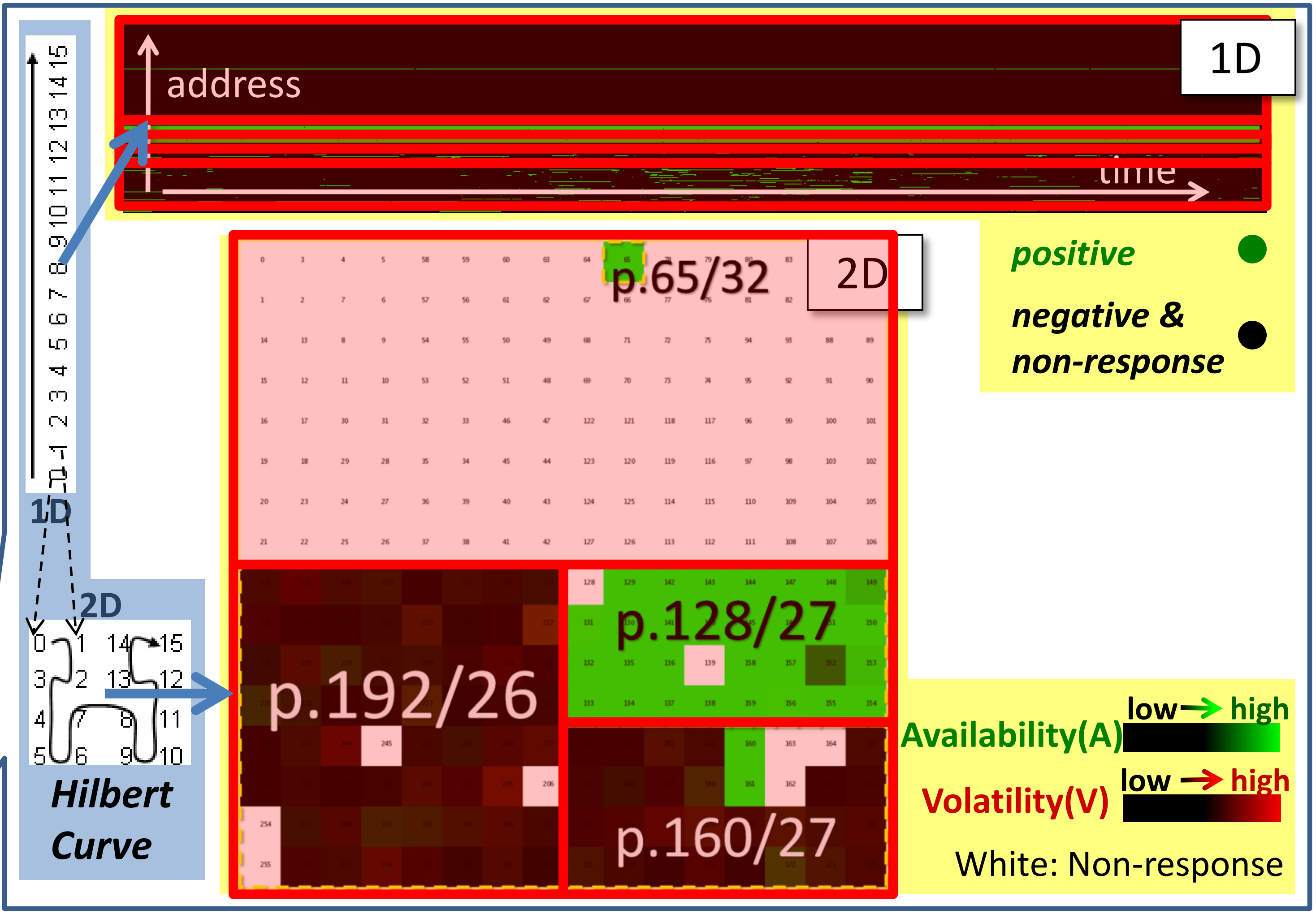
Example:
= median(2,2,1) = 2

Intuition:
typical duration

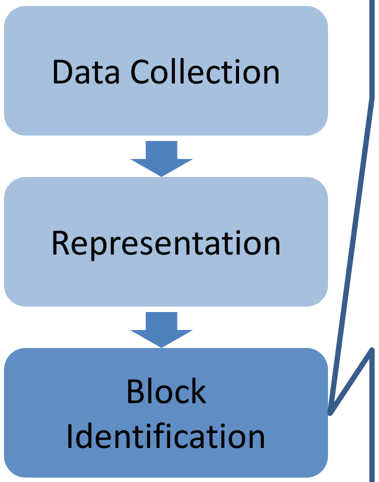


Spatial Correlation: Block Identification





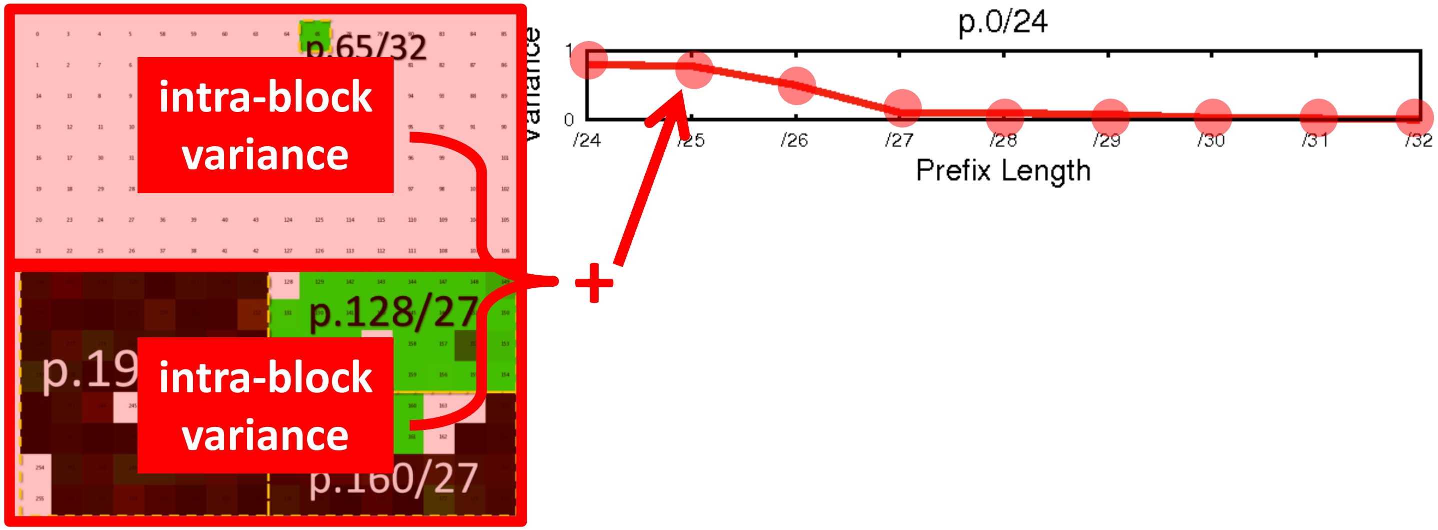
Spatial Correlation: Block Identification



Data Collection

Representation

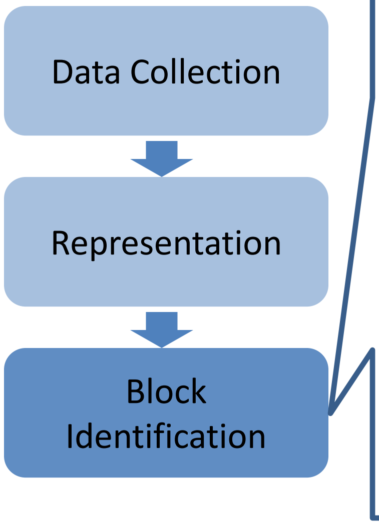
Block Identification

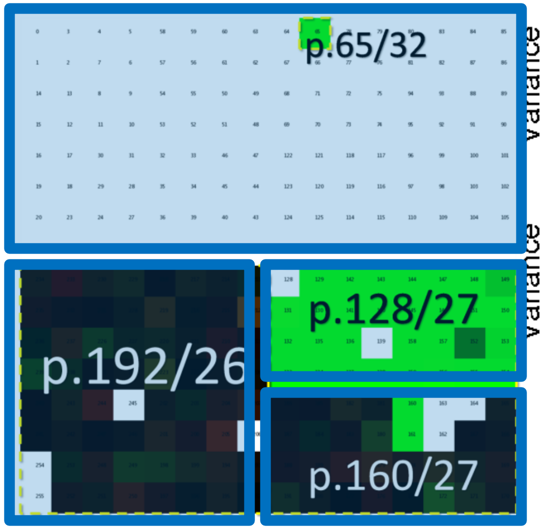


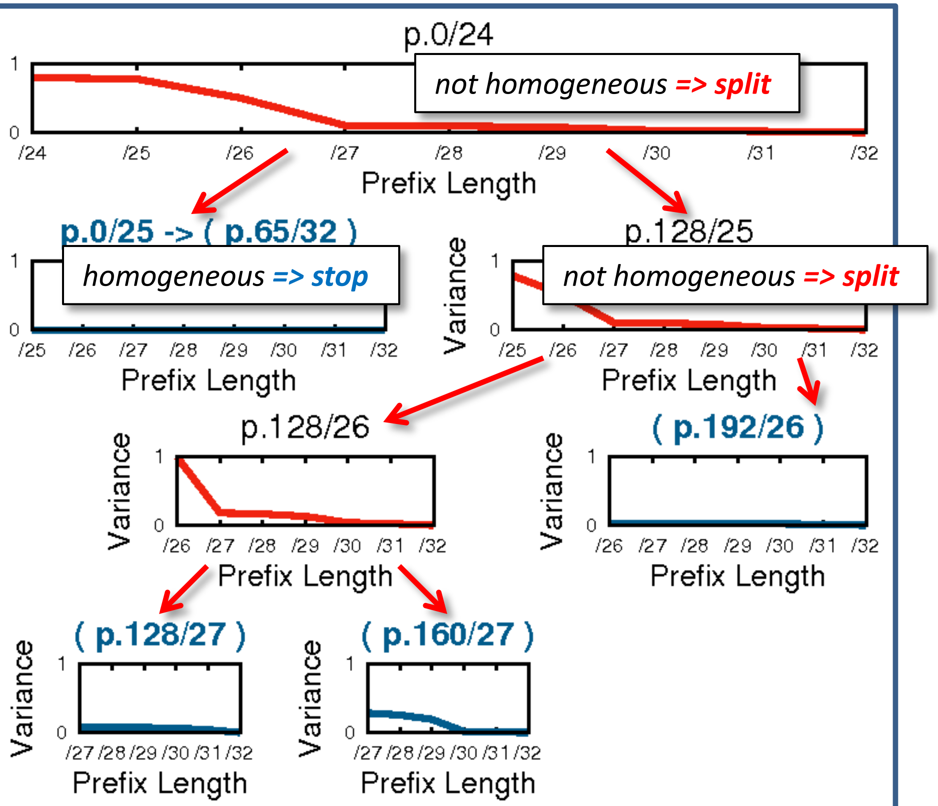
Idea: **examine each block size, if block is homogeneous, stop else split and recurse**

How

Spatial Correlation: Block Identification







Idea: **examine each block size, if block is homogeneous, stop else split and recurse**

How

Spatial Correlation: Validation

- Validation is **hard**
  - Where to find ground truth?
    - decentralized management
    - usage block ground truth?

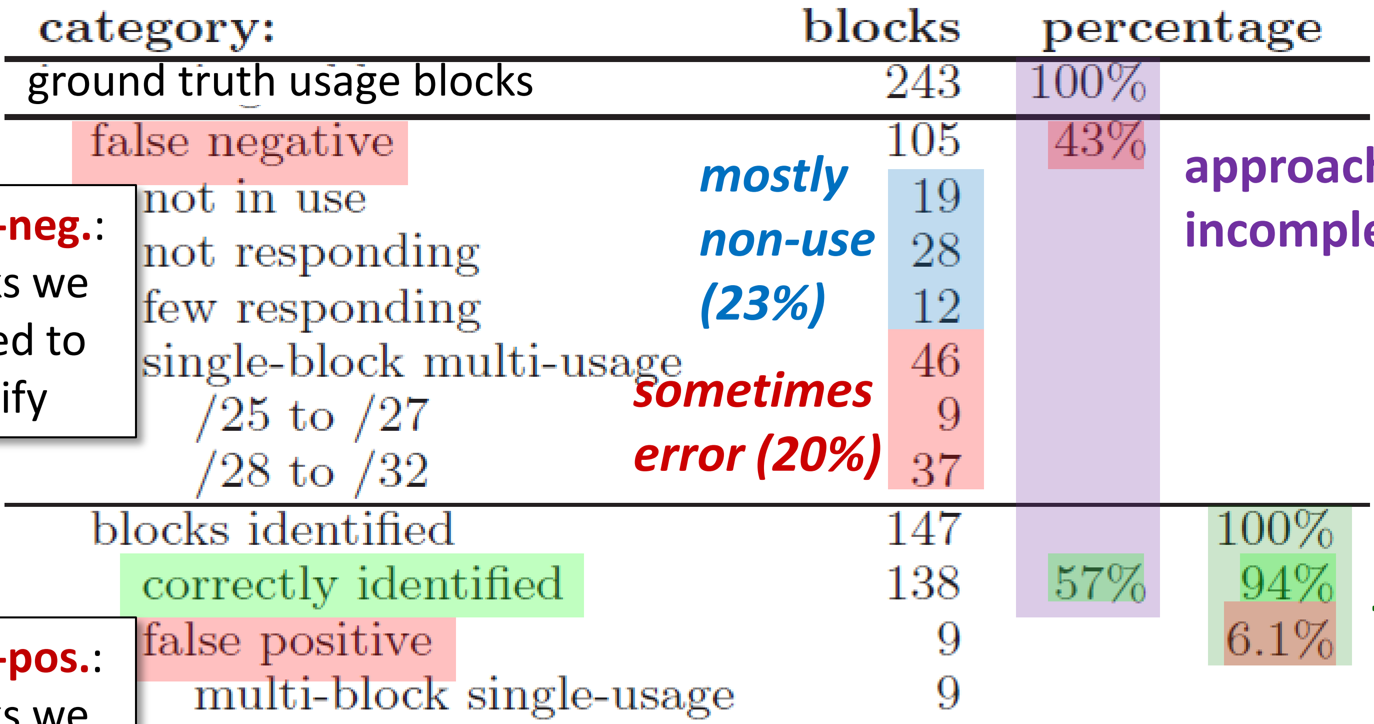
- Use three complementary ways:
  - Compare to USC’s network (*operator provided truth*)
  - Compare to general Internet (*hostname inferred truth*)
  - Evaluate different samples and dates
    - is 1% of the Internet enough? yes!
    - trends change some over time
    - details: paper section 5.3

Spatial Correlation: USC’s Network

- **Why**
  - *quite solid truth (operator provided)*
  - knowledge of both ***allocated blocks*** and ***usage blocks***

- **How**
  - compare ***observable blocks*** (result to validate) with ***usage blocks*** (ground truth)

Spatial Correlation: USC's Network



category:	blocks	percentage	
ground truth usage blocks	243	100%	
false negative	105	43%	
not in use	19		
not responding	28		
few responding	12		
single-block multi-usage	46		
/25 to /27	9		
/28 to /32	37		
blocks identified	147		100%
correctly identified	138	57%	94%
false positive	9		6.1%
multi-block single-usage	9		

mostly non-use (23%)

sometimes error (20%)

approach is incomplete

but what is found is correct

false-neg.: blocks we missed to identify

false-pos.: blocks we wrongly identified

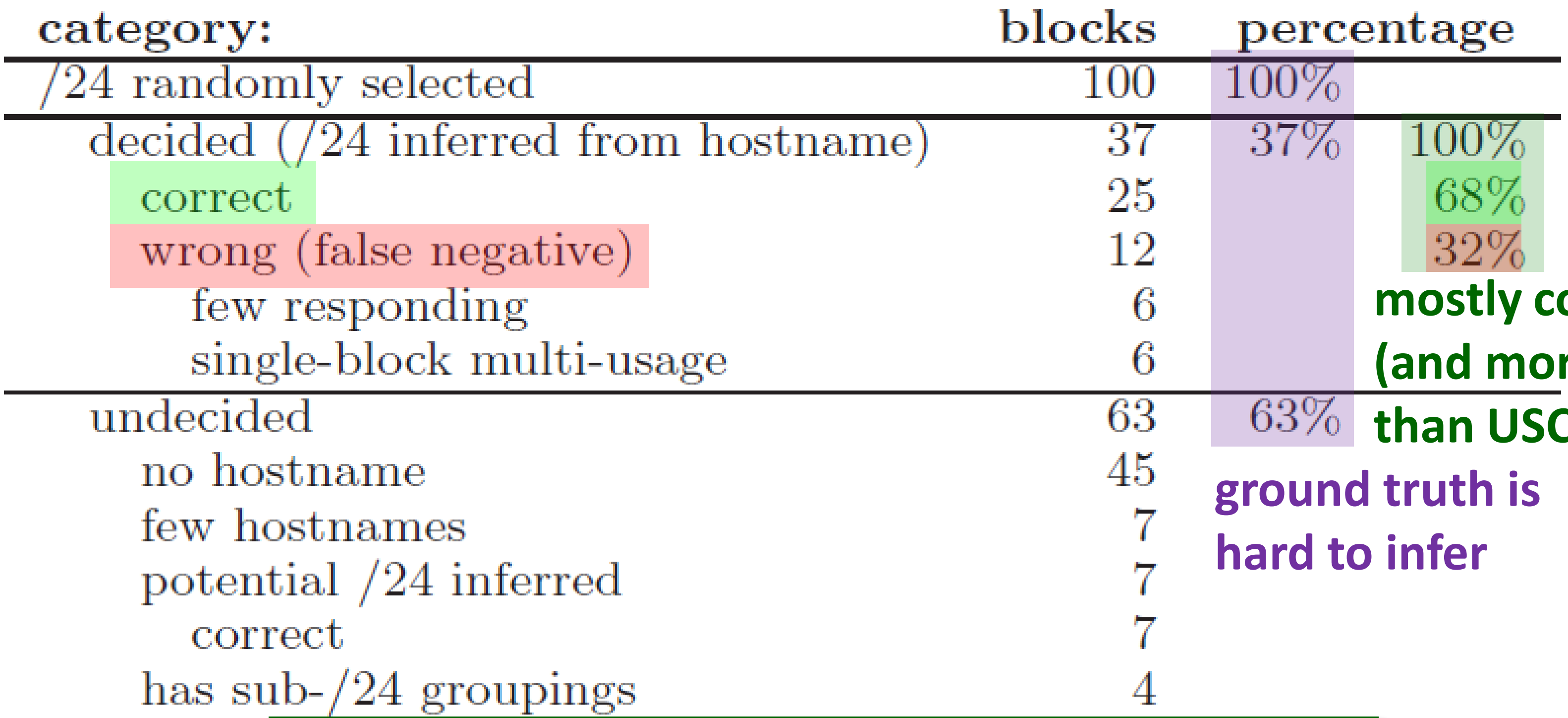
very accurate when it reaches a conclusion

Spatial Correlation: General Internet

- **Why**
  - *unbiased truth (randomly selected)*

- **How**
  - Infer ***usage blocks*** from hostnames
    - dhcp-host-xxx.example.net
  - compare ***observable blocks*** (result to validate) with ***usage blocks*** (ground truth)

Spatial Correlation: General Internet



category:	blocks	percentage	
/24 randomly selected	100	100%	
decided (/24 inferred from hostname)	37	37%	100%
correct	25		68%
wrong (false negative)	12		32%
few responding	6		
single-block multi-usage	6		
undecided	63	63%	
no hostname	45		
few hostnames	7		
potential /24 inferred	7		
correct	7		
has sub-/24 groupings	4		

mostly correct (and more than USC)

ground truth is hard to infer

methodology more complete when evaluate with unbiased sample

Key Contributions

	Methodology	Application	Validation
Spatial Correlation?	Group addresses into blocks by usage	More frequent probing? Block sizes? Block-level usage?	USC’s network, General Internet, Consistency
Address Utilization?	**See paper**	**See paper**	**See paper**
Dynamic Addressing?			
Low-bitrate Identification?	Utilize standard deviation of RTTs	Auto content serving? Network management?	USC’s network, General Internet

Background: What is low-bitrate?

- Addresses are connected to Internet through edge access links
- Different access link type has different bitrate
  - *Dial-up*: 56Kb/s
  - *ADSL* (typical): 3,000/768 kbit/s
  - *GPRS*: 57.6 Kb/s
  - *UMTS 3G*: 384 kbit/s

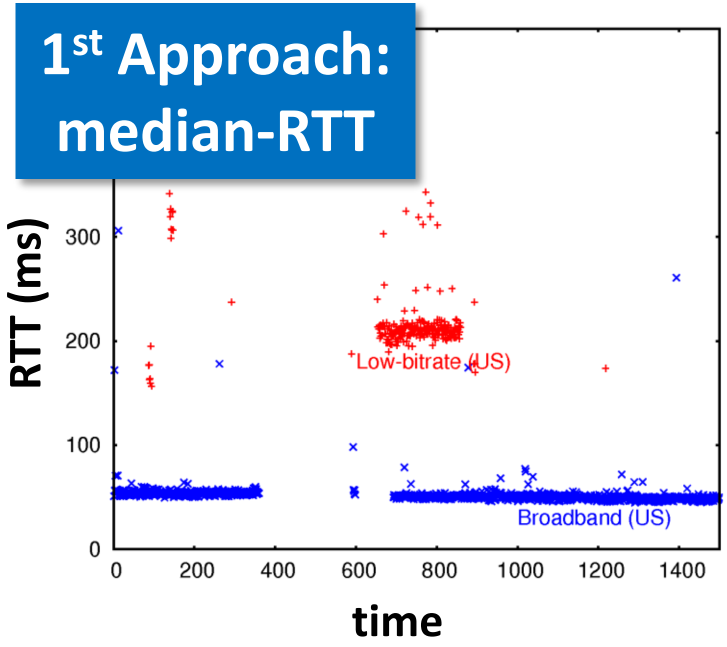
- **We define low-bitrate as less than 100Kb/s**, such as dial-up and GPRS.

Low-bitrate: Application

- Why care?
  - For the researchers
    - help understand trends in technology deployment
  - For the business
    - automatically match content and layout
  - For network management
    - low-bitrates links are correlated with short connect-times and sparse usage.

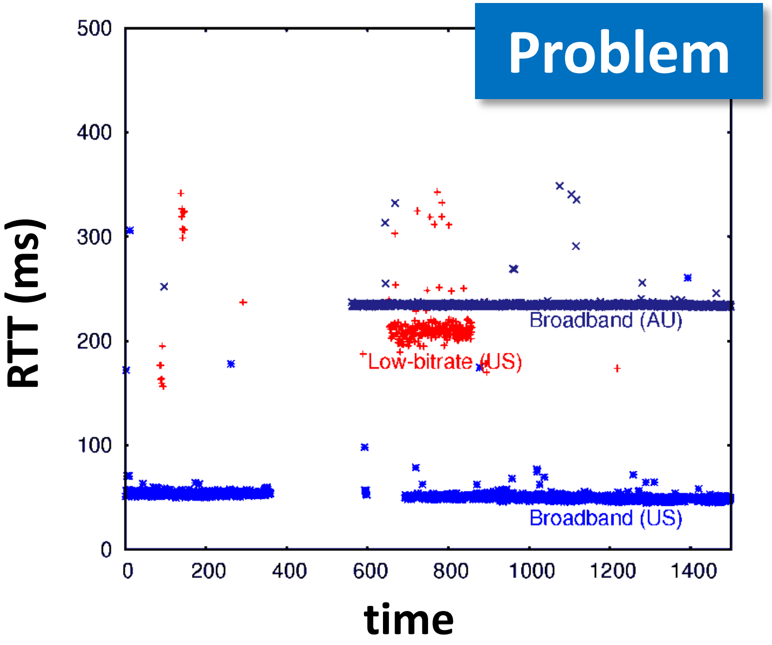
Methodology: Formalizing RTT -> Edge Bitrate

- RTT = **transfer** + queuing + ***propagation***



1st Approach: median-RTT

transfer distinguishes low-bitrate vs. broadband



Problem

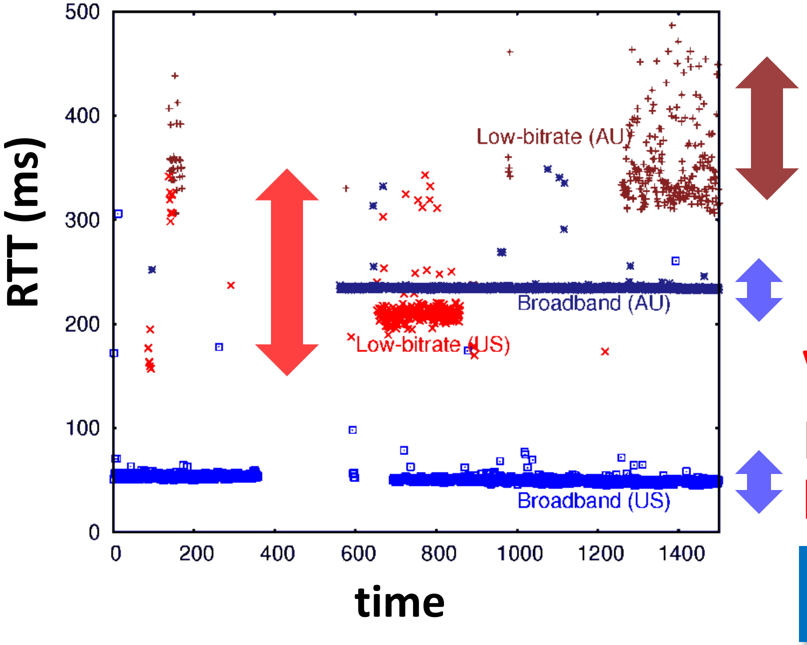
but internationally ***propagation** time dominates*

Methodology: Formalizing RTT -> Edge Bitrate

- RTT = transfer + **queuing** + propagation

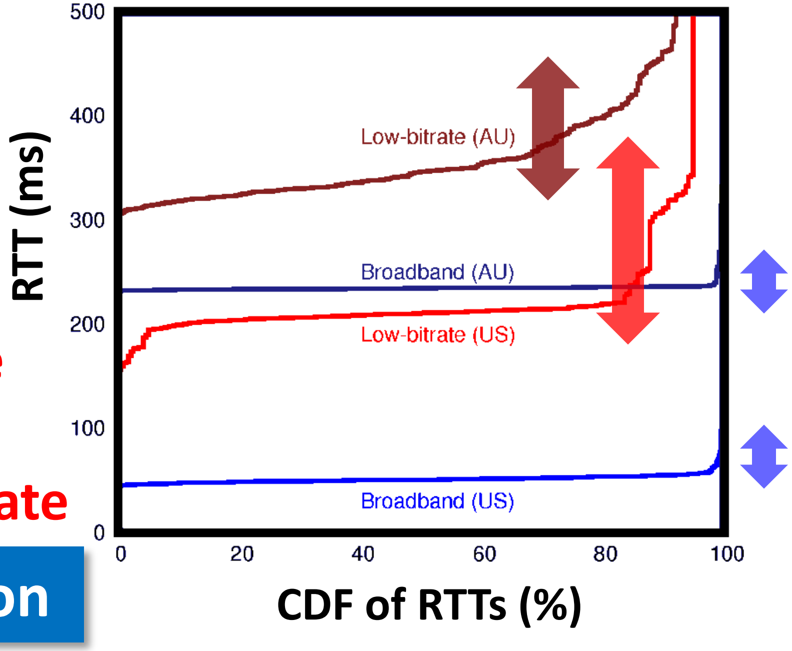
edge-bitrate dependent, and ***varying***

distance dependent, but ***consistent***



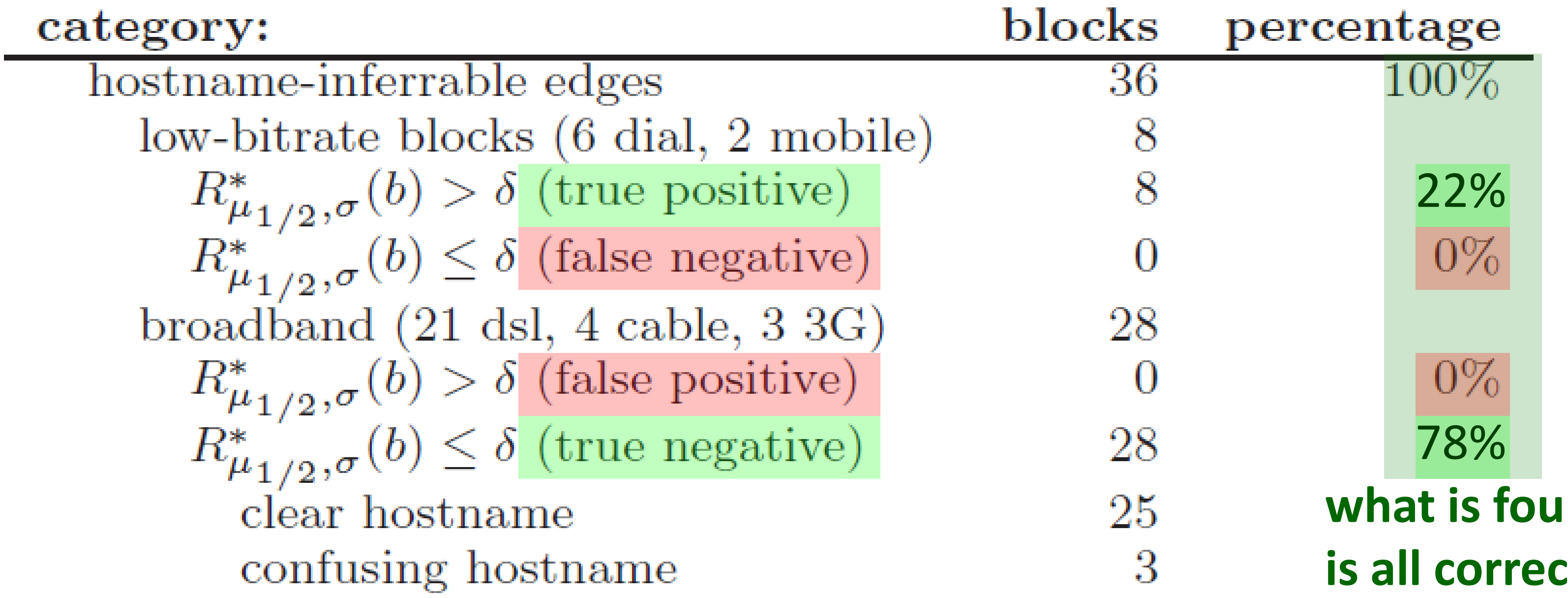
variance predicts **low-bitrate**

Solution



(or consistency predicts **broadband)**

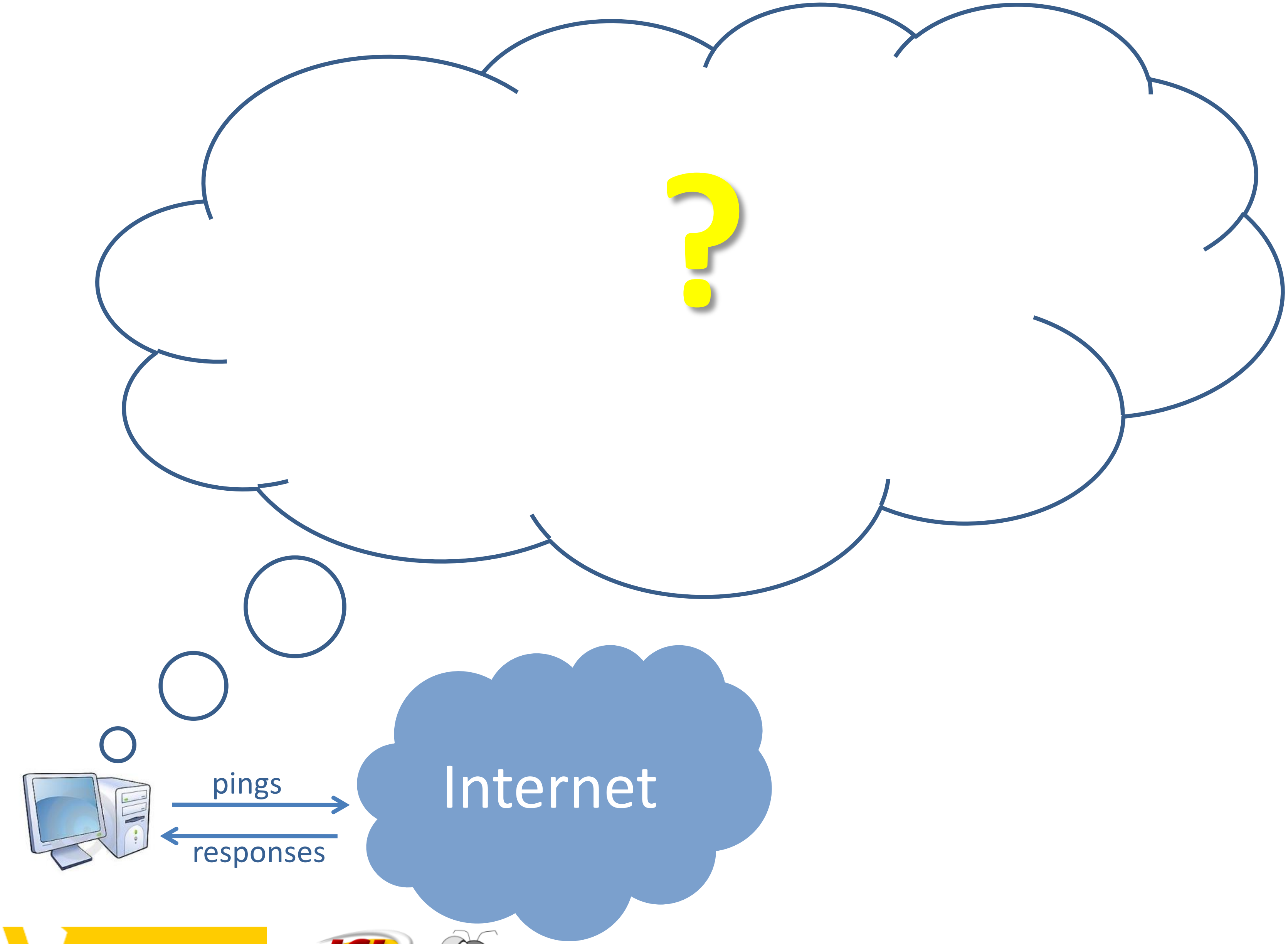
Low-bitrate: Validation



category:	blocks	percentage
hostname-inferrable edges	36	100%
low-bitrate blocks (6 dial, 2 mobile)	8	
$R^*_{\mu_{1/2},\sigma}(b) > \delta$ (true positive)	8	22%
$R^*_{\mu_{1/2},\sigma}(b) \leq \delta$ (false negative)	0	0%
broadband (21 dsl, 4 cable, 3 3G)	28	
$R^*_{\mu_{1/2},\sigma}(b) > \delta$ (false positive)	0	0%
$R^*_{\mu_{1/2},\sigma}(b) \leq \delta$ (true negative)	28	78%
clear hostname	25	
confusing hostname	3	

what is found is all correct

can accurately find low-bitrate links



?

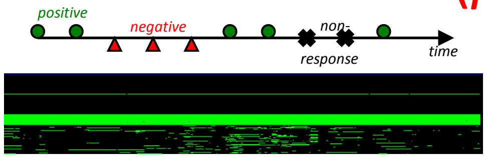
pings

responses

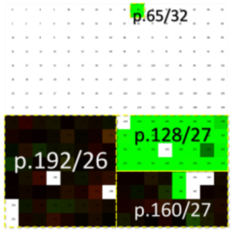
Internet

Conclusion

SIMPLE observations (*pings*)



can tell …

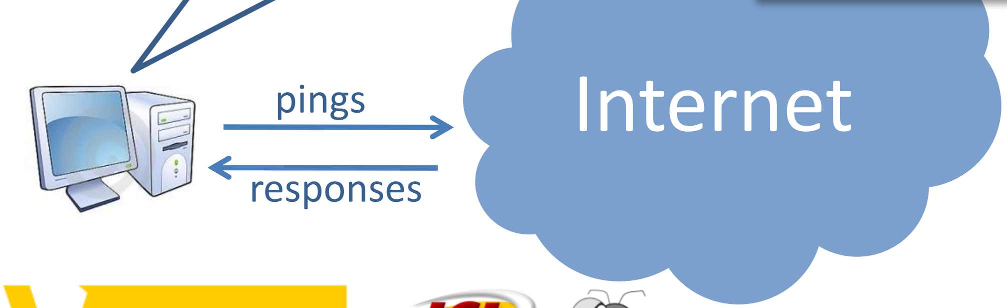


VALUABLE truths about the Internet.

spatial correlation, address utilization

dynamic addressing, low-bitrate

Visit www.isi.edu/ant
for our dataset and more information!


pings
responses
Internet