Prologue

Yuval Shavitt
School of Electrical Engineering
TEL AVIV UNIVERSITY אוניברסיטת תל-אביב
shavitt@eng.tau.ac.il

http://www.netDimes.org
http://www.eng.tau.ac.il/~shavitt

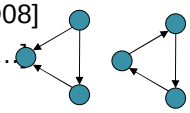
DIMES: Why and What

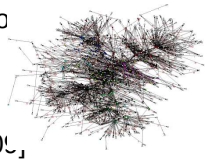
- Diminishing return?
  - Replace instrumentation boxes with software agents
  - Ask for volunteers do help with the measurement

    ⇓
  - The cost of the first agent is very high
  - each additional agent costs almost zero
- Advantages
  - Large scale distribution: view the Internet from everywhere
  - Remove the “academic bias”, measure the commercial Internet
- Capabilities
  - Anything you can write in Java!
  - Obtaining Internet maps at all granularity level with annotations
    - connectivity, delay, loss, bandwidth, capacity, jitter, ….
  - Tracking the Internet evolution in time
  - Monitoring the Internet in real time

What do we do with DIMES?

- DIMES data analysis
  - *k*-shell analysis [Carmi *et al*., PNAS07]
  - Bias analysis [Weinsberg & S., Infocom 09; …]
  - Anonymous router identification [Almog *et al*., MCD08]
  - Efficient motif identification [Gonen & S., WAW09; …]
- Generating periodic PoP level maps
  - Coarse PoP identification [Feldman & S., Globeco
  - PoP Geo-location [under work]
- New Measurements
  - Packet Trains [Allalouf, Kaplan & S., Tridentcom0
  - ParisTraceroute
- Optimizing DIMES operation
  - Approximation results [Gonen & S., IPL 09; …]





DIMES and You

- Data is available to all
  - Periodic topologies are on the web
  - Other data is gladly shared by request
- Others are running distributed experiments thru Web
  - easy to use
- Easy to add new capabilities
- Future
  - Open DIMES data for applications
    - Internet distance service
    - Improve P2P application
  - PlanetLab deployment (within days)
- We can also use your help: download an agent
  http://www.netDimes.org

Other measurement activities

- P2P Networks
  - 15-40% of queries to Gnutela for >100 days
    - Spatial-temporal analysis of Gnutela queries [Gish, Tankel, S., IPTPS'07]
    - Predicting artist success from queries [Koenigstein, S., Tankel, KDD'08; …]
  - Disk content for 1.2M users in same day
    - Content clustering [Weinsberg, Weinsberg, S, submitted]
  - DC queries collection effort
- Cellphone network
  - 1 Million private users. monthly summaries of calls, talk time, SMSs
  - Data on users: age, gender, zip, group
  - Commercial data

Quantifying the Importance of Vantage Points Distribution in Internet Topology Measurements

Yuval Shavitt and Udi Weinsberg
School of Electrical Engineering
Tel-Aviv University
Israel

Goals

- Bias
  - Does the distance from the measurement vantage points (VPs) skew our topology characteristics?
- Quantify the importance of a *diverse* and *broad* set of VPs on the resulting topology.

Data Set

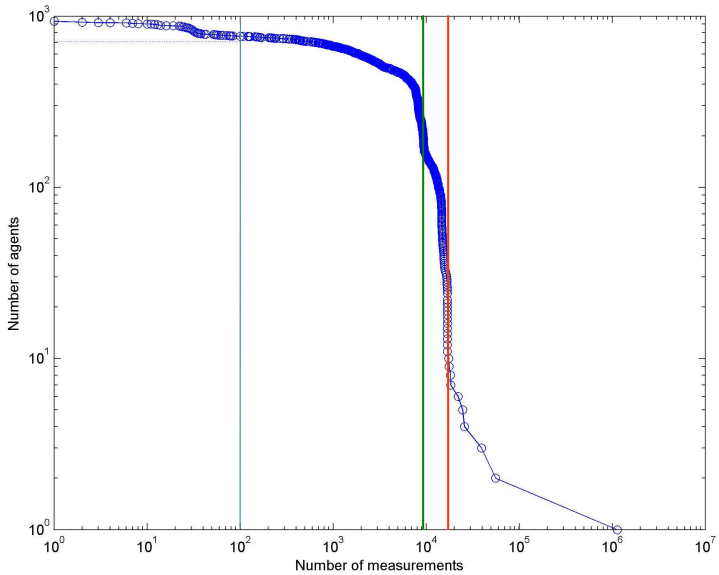
- Data is obtained from DIMES
  - Community-based infrastructure, using almost 1000 active measuring software agents
  - Agents follow a script and perform ~2 probes per minute (ICMP/UDP traceroute, ping)
  - Most agents measure from a single AS (vp)
    - But some (appear to) measure from more…
    - Data need to be filtered to remove artifacts
  - Traceroute data collected during March

Filtering the data

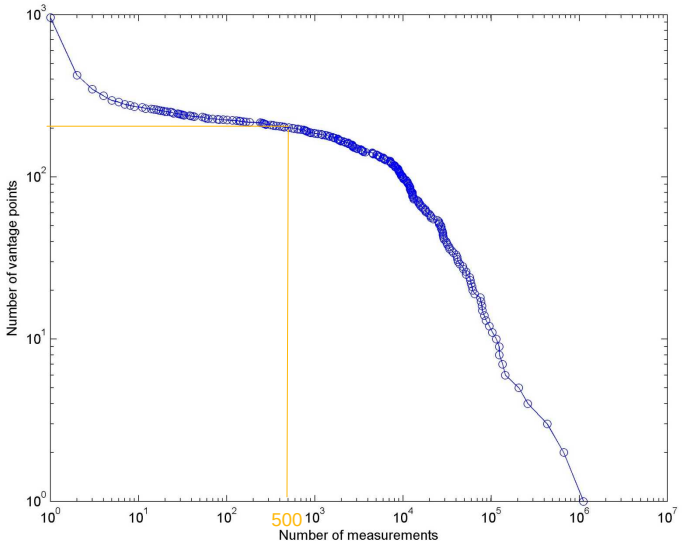
- For each agent and each week, classify how many networks it measured the Internet from

  Typical cases:
  - AS_i:15300, AS_j:8
  - AS_i:10000, AS_j:3178
  - AS_i:10000, AS_j:412 , AS_k:201
  - 18000, 12, 11, 9, 9, 3, 3, 2, 2, 1, 1, 1, 1, 1, ….

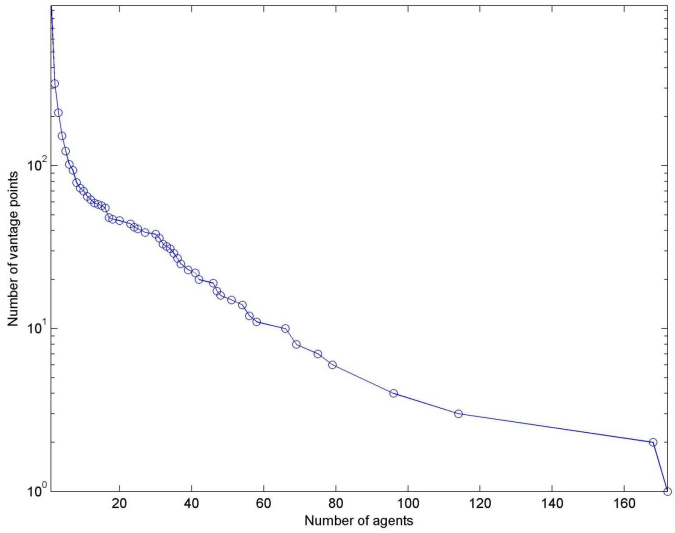
Measurements Per Agent



Measurements per Network

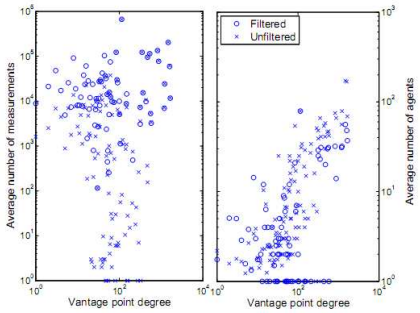


Agents per Network



Filtering Results

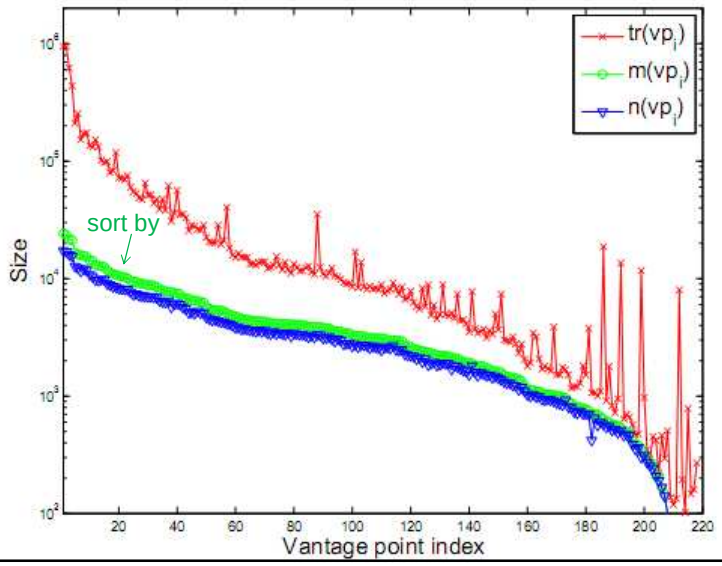
- 96% of the agents have less than 4 different vps
- High degree ASs tend to have more agents
- High number vps degrees



Diminishing Returns?

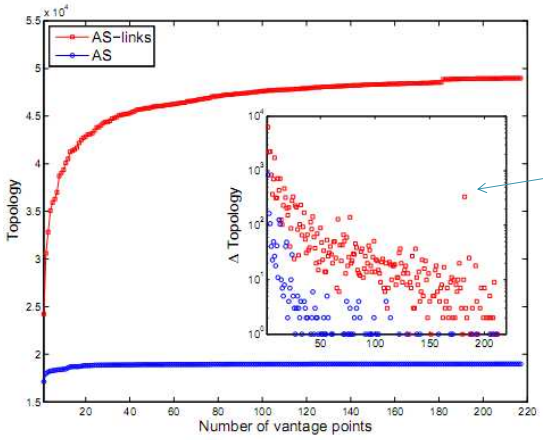
- Barford *et. al.* – the utility of adding many vps quickly diminishes
  - In terms of ASes and AS-links
- Shavitt and Shir – utility indeed diminishes but the tail is long and significant
  - Tail is biased towards horizontal links
- We wish to quantify how different aspects of AS-level topology are affected by adding more vps

Creating topologies per VP



Topology Size

- The return (especially for AS links) does not diminishes fast!

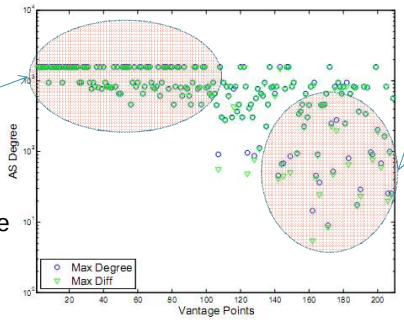


VP with small local topology can contribute many new links!

Direction of Detected Links

- For each link: Plot max adjacent AS degree and max adjacent ASes degree difference

High degree difference – indicates radial links towards the core



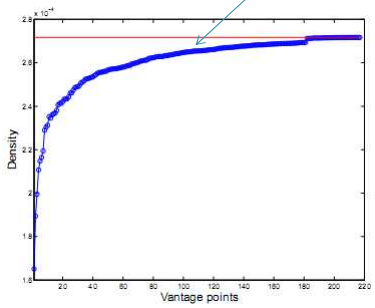
Low degree difference – indicates tangential links and links between small-size ASes

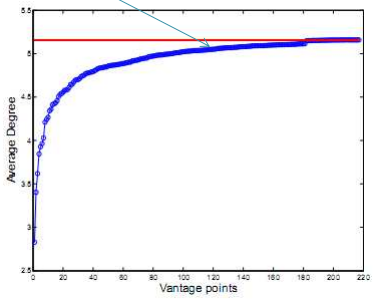
Convergence of Properties

- Taking several common AS-level graph properties, and analyze their convergence as local topologies are added
  - Keeping the sort order by number of links
- Slow convergence indicates the need to have broad and diverse set of vps

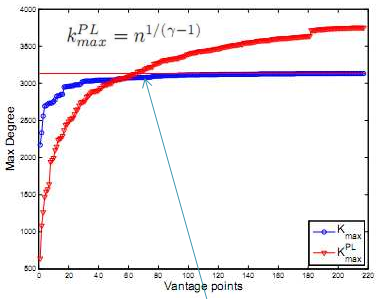
Density and Average Degree

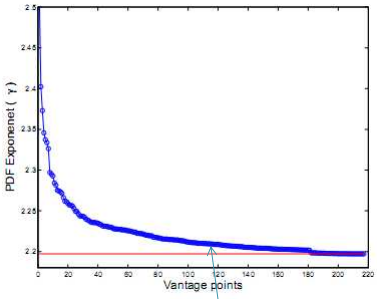
Slow convergence of density and average degree – easy to detect ASes but difficult to find all links





Power-law and Max Degree

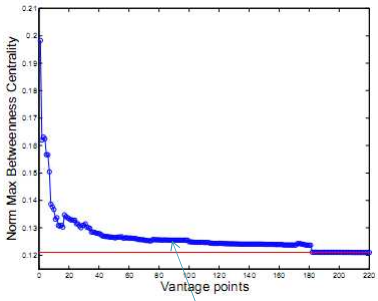




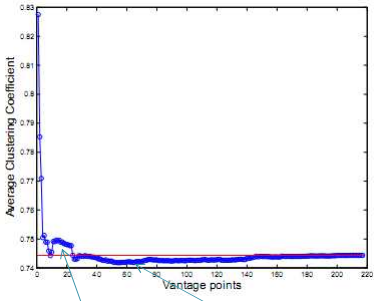
Fast convergence of maximal degree – core links are easily detects

Fair convergence of power-law exp$\gamma = 2.197$

Betweenness and Clustering



Fast convergence of max bc – Level3 (AS3356), a tier-1 AS is immediately detected as having max bc



Radial links decrease cc

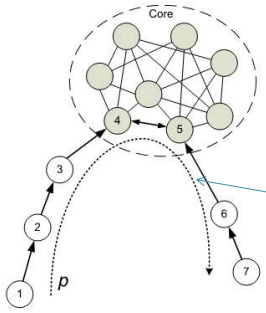
Tangential links increase cc

Revisitng Sampling Bias

- Lakhina *et al.* – AS degrees inferred from traceroute sampling are biased
  - ASes in vicinity to vps have higher degrees
  - Power-law might be an artifact of this!
- Dall'asta *et al.* – no…it is quite possible to have unbiased degrees with traceroutes
- Cohen *et al.* – when exponent is larger than 2, resulting bias is neglible

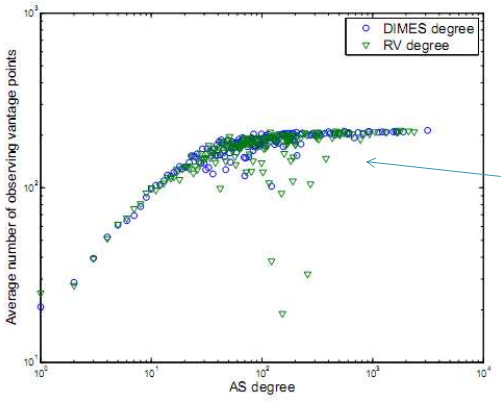
Evaluating Sampling Bias

- For each AS find:
  - All the vps that have it in their local topology
  - The Valley Free distance in hops



Up-hill to the core (c2p), side-ways in the core (p2p) and down-hill from the core (p2c)

Dataset VPs and Distances



Low degree ASes are seen from less vps than high-degree Ases…this makes sense!

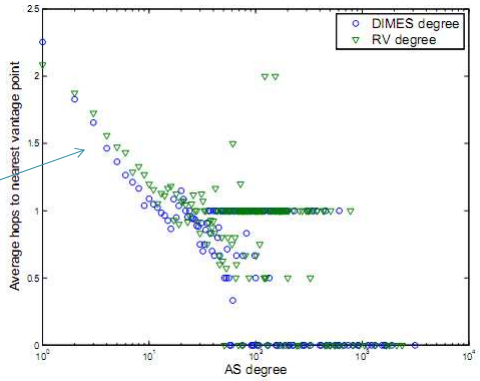
In our dataset, most ASes have a vp that is only 1-2 hops away!

DISTANCES FROM AS TO NEAREST VANTAGE POINT

Distance to nearest vp	0	1	2	3	4	5
Number of ASes	213	6365	9329	2869	204	6

Average Distance per Degree

Low degree ASes are seen from farther vps…sampling bias?



No real bias!

- More VPs are located in high-degree ASes
- There are high-degree ASes that are seen from "far" vps
- Broad distribution – all ASes are pretty close-by to a vp!

Revisiting Diversity Bias

- What is the effect of diversity in vps geo-location and network type?
  - Some infrastructures rely on academic networks for vp distribution – does it have an effect on the resulting topology?
- We compare iPlane and DIMES
  - Classify AS into types: t1,t2, edu, comp, ix, nic using Dimitropoulos *et al.*

Diversity Bias Evaluation

VANTAGE POINTS' TYPES OF IPLANE AND DIMES

	t1	t2	edu	comp	ix	nic	unknown
iPlane	17	104	117	22	3	5	46
DIMES	23	129	6	15	0	0	47

iPlane uses many PlanetLab nodes (edu), while DIMES resides mostly at homes (tier-2)

ASES' TYPES FOR WHICH DIMES DEGREES ARE LARGER THAN IPLANE'S (RATIO<1), SMALLER (RATIO>1) AND EQUAL (RATIO=1)

	t1	t2	edu	comp	ix	nic	unknown
ratio<1	20	1487	108	1398	7	34	1286
ratio>1	24	1295	137	1219	9	36	1123
ratio=1	0	1401	248	3304	11	61	2445

Indeed DIMES have higher t2 and comp degrees and iPlane have higher edu degrees – results are slightly biased to vps' types!

In Search of Ground Truth

- One week is not sufficient for active measurements

AVERAGE AS DEGREES USING ROUTEVIEWS, DIMES AND IPLANE PER AS TYPES

	t1	t2	edu	comp	ix	nic	unknown
RouteViews	575.84	9.1	3.4	2.29	14.85	5.42	3.2
DIMES	528.31	8.36	3.34	2.3	7.96	4.06	3.86
iPlane	565.31	8.47	5.83	2.45	22.82	4.12	4.5

- Both iPlane and DIMES have lower average degrees than RouteViews
  - Except iPlane's edu and ix!
  - Diversity bias exists – need diverse vp types!

Measuring Within a Network

- Comparing vp average degrees to quantify the effect of measuring within a network

VPs	Avg. Degree	t1	t2	edu	comp	ix	nic
iPlane	iPlane	951.5	36.0	9.4	8.1	172.0	3.5
	DIMES	961.3	27.0	5.1	5.5	28.6	3.5
DIMES	iPlane	904.7	38.0	2.6	9.1	-	-
	DIMES	903.1	50.1	18.6	25.8	-	-
Both	iPlane	1157.0	89.5	3.0	2.0	-	-
	DIMES	1161.0	81.3	14.0	3.0	-	-

Indeed, the average degree when measuring within a network is mostly higher (hmm…tier-1 doesn't count cause most vps are the same!)

Conclusion

- VP distribution is important
  - Number, AS type, geo-location
- AS-level graph properties are affected
  - Some converge very fast
  - Other converge slowly
- Community based projects have practically unlimited growth potential!