The Joint Degree Distribution as a Definitive Metric of the Internet AS-level Topologies

Priya Mahadevan, Dimitri Krioukov, Marina Fomenkov,
Brad Huffaker, Xenofontas Dimitropoulos, kc claffy,
Amin Vahdat

CAIDA, SDSC, UC San Diego

Published in ACM SIGCOMM CCR, January 2006
www.caida.org/publications/papers/2006/as_topology

2006 ISMA WIT, San Diego, May 2006

Plan

- Data sources:
  - collection methodologies
  - limitations
- Graph metrics
  - definitions
  - values in our graphs
  - interdependencies
- Joint Degree Distribution (JDD) - the definitive metric
  - defines values of other metrics
  - captures crucial graph properties
- Comparison of observed graphs with random graph models

Data Sources - 1

- BGP Tables
  Border Gateway Protocol - for routing among ASes
- RouteViews collects BGP routing tables
  www.routeviews.org
  - 7 collectors, each has a number of globally placed peers
  - archives both static snapshots and dynamic data
  - data are publicly available
- For this study - data from March 2004
  - used collector with the largest # of peers = 68
  - discarded AS-sets and private ASes
  - merged 31 daily graphs into one graph

$=>$ *BGP graph*

Data Sources - 2

- Traceroute
  sequence of IP hops along the forward path from the
  to a given destinations
- CAIDA traceroute-based tool *skitter*
  - continuous measurements since 1998
  - more than 20 monitors all over the world
  - destination list of about a million IPv4 addresses
- For this study - data from March 2004
  - mapped IP addresses to origin AS numbers
    using BGP tables from RouteViews
  - discarded about 5% of links
    ambiguous mappings, measurement inaccuracies
  - merged 31 daily graphs into one graph

  $=>$ *skitter graph*
- daily derived AS-level topology graphs available at
  www.caida.org//tools/measurement/skitter/as_adjacencies.xml

Data Sources - 3

- WHOIS
  a collection of databases with AS peering information
  - manually maintained
  - no timely updates
  - *RIPE WHOIS* is the most current and reliable
    (but covers mostly European infrastructure)
- For this study -

  RIPE WHOIS database dump, April 7, 2004
  - looked for records indicating links btw ASes
  - discarded external and private ASes

  $\Rightarrow$ *WHOIS graph*

Data Sources (cont.)

The three graphs present different views of the Internet

- skitter graph
  - topology of actual Internet traffic flows
  => *data plane*
- BGP graph
  - topology of the routing system
  => *control plane*
- WHOIS graph
  - topology created by human actions
  => *management plane*

- both skitter and BGP are *traceroute-like* explorations
- WHOIS reports peering arrangements made by humans
- we verified that differences betwen WHOIS and the other two graphs are not due to geographical bias

Topology Characteristics

Average Degree

- n = number of nodes (or *graph size*)
  m = number of links
- average node degree $\bar{k} = 2m/n$
- the coarsest connectivity characteristic

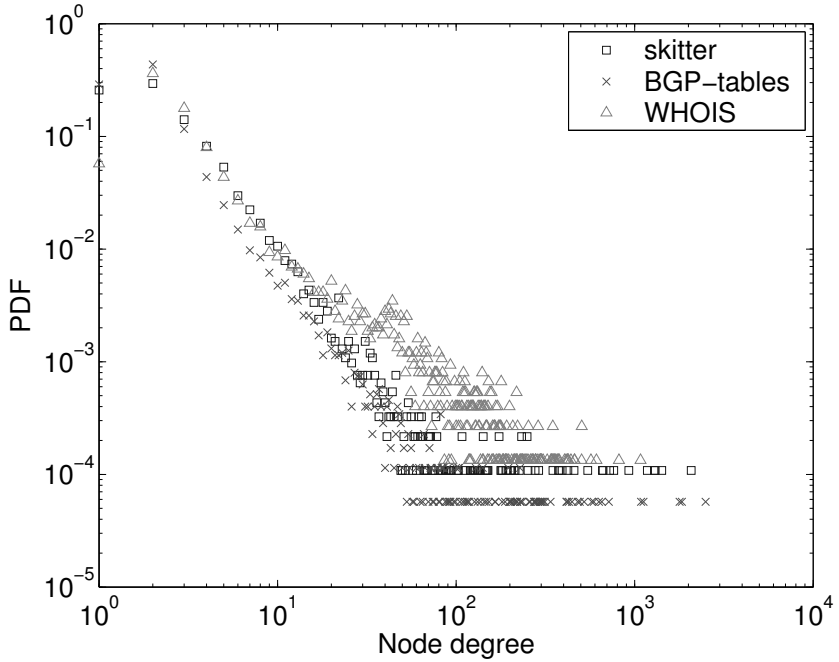
	skitter	BGP tables	WHOIS
Number of nodes (n)	9,204	17,446	7,485
Number of edges (m)	28,959	40,805	56,949
Avg node degree $(\bar{k})$	6.29	4.68	15.22

- **$\bar{k}$-order:** BGP - skitter - WHOIS
  increasing average degree

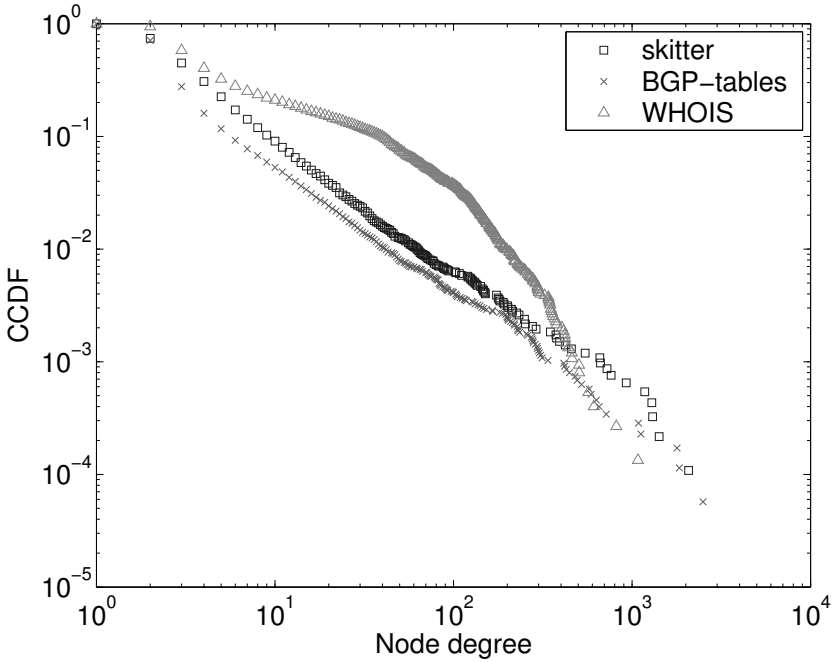
Degree Distribution

- $n(k)$ = number of nodes of degree k (k-degree nodes)
- $P(k) = n(k)/n$ – probability that a node is k-degree

PDF



CCDF



- both PDFs and CCDFs are in the $\bar{k}$-order, BGP-skitter-WHOIS
- skitter graph is closest to power law, $\gamma = -2.25$
- WHOIS graph is not power law at all
  has excess of medium-degree nodes

Joint Degree Distribution (JDD)

- $m(k_1, k_2)$ = number of edges connecting nodes of degrees k_1 and k_2
- $P(k_1, k_2) = \mu(k_1, k_2) \times m(k_1, k_2)/(2m)$
  (where $\mu(k_1, k_2)$ is 1 if $k_1 = k_2$ and 2 otherwise)
  - probability that an edge connects k_1- and k_2-degree nodes
- JDD contains more information about connectivity in a graph than degree distribution
- JDD provides information about 1-hop neighborhoods around a node
- given JDD $P(k_1, k_2)$, one can always restore $P(k)$ and $\bar{k}$

Joint Degree Distribution (JDD) - cont.

- summary statistic of JDD: assortativity coefficient

  $r \sim \sum_{k_1,k_2=1}^{k_{max}} k_1 k_2 (P(k_1,k_2) - k_1 k_2 P(k_1) P(k_2)/\bar{k}^2)$

  $-1 \leq r \leq 1$

  – directly related to *likelihood* defined by Li *et al.*

- *disassortative* networks with $r < 0$:
  - excess of *radial* links connecting nodes of dissimilar degrees
  - vulnerable to random failures and targeted attacks
- *assortative* networks with $r > 0$:
  - excess of *tangential* links connecting nodes of similar degrees

Joint Degree Distribution (JDD) - cont.

- all three our graphs are disassortative

	skitter	BGP tables	WHOIS
Assortativity Coefficient (r)	-0.24	-0.19	-0.04

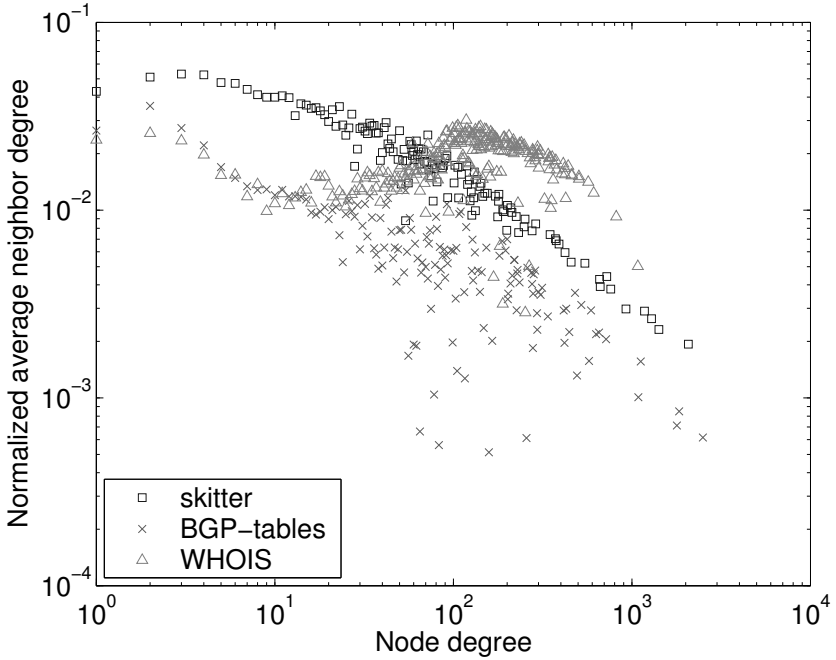
- **r-order:** WHOIS - BGP - skitter
  decreasing assortativity coefficient
- both skitter and BGP are *traceroute-like* explorations
  - discover more radial links
    (connecting low-degree customer ASes and high-degree large ISP ASes)
  - fail to detect tangential links
    (connecting nodes of similar degrees)
- WHOIS-based methodology finds abundant medium-degree tangential links
  $=>$ WHOIS graph is more assortative

Joint Degree Distribution (JDD) - cont.

- summary statistic of JDD:

  the average neighbor connectivity $k_{nn}(k) = \sum_{k'=1}^{k_{max}} k' P(k'|k)$

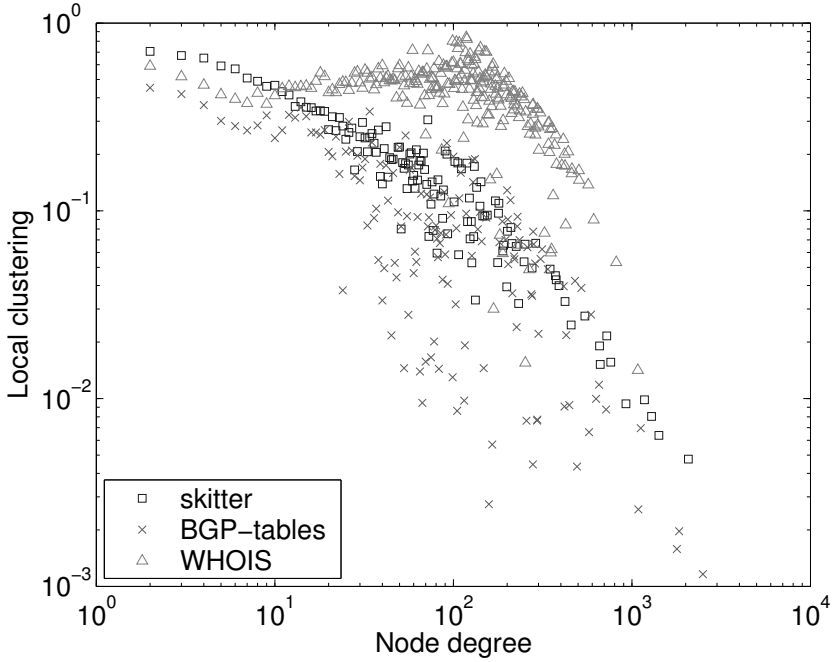
– the average neighbor degree of the average k-degree node



- low degrees – r-order, skitter at the top
- high degrees – $\bar{k}$-order, WHOIS at the top

Clustering

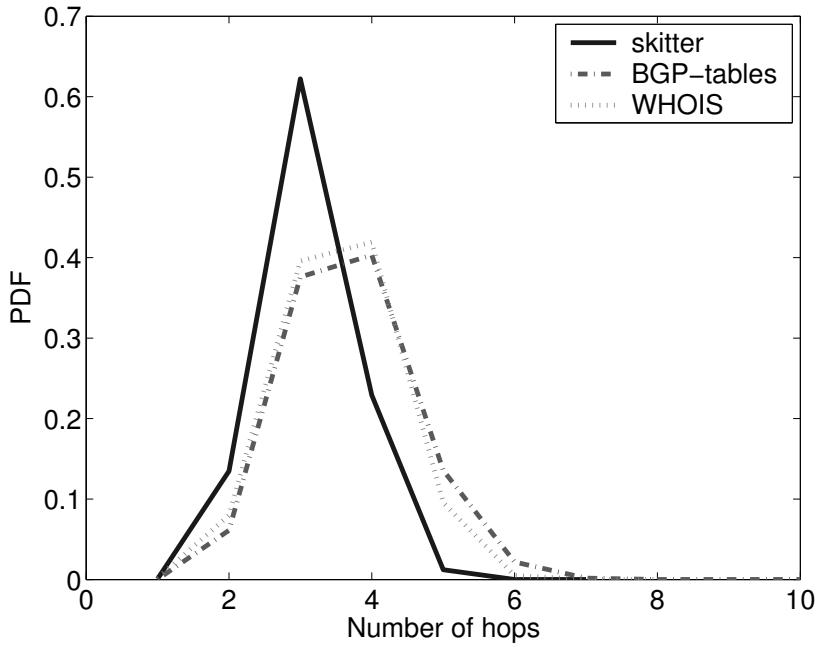
- $\bar{m}_{nn}(k)$ = number of links between the neighbors of k-degree nodes
  $k(k-1)/2$ = the maximum possible number of such links
- $C(k) = 2\bar{m}_{nn}(k)/[k(k-1)]$ – local clustering



- low degrees – r-order, skitter at the top
- high degrees – $\bar{k}$-order, WHOIS at the top

Distance

- $d(x)$ - distance distribution, the probability that two random nodes are at a distance x hops from each other



- interplay between $\bar{k}$-order and r-order
  - skitter - most disassortative $=>$ shortest average distance
  - BGP - less dense, lower $\bar{k}$ $=>$ larger average distance
  - WHOIS - more assortative, higher r $=>$ larger average distance

Topology Characteristics

- other topology metrics:
  - rich club connectivity
  - coreness
  - eccentricity
  - betweenness
  - spectrum

  www.caida.org/analysis/topology/as_topo_comparisons/
- statistics tables, plots, and calculated data used to draw them
- metric values and differences in the three graphs can be explained using $\bar{k}$-order and r-order

Comparison with random graph models

- Random graph models:
  - 0K – reproduces average degree $\bar{k}$
  - 1K – reproduces degree distribution $P(k)$
    - power-law random graphs (PLRG)
  - 2K – reproduces JDD $P(k_1, k_2)$
- 0K- and 1K-random graphs are *uncorrelated*
  (when forced correlations are not taken into account)
  - assortativity coefficient $r = 0$
  - the average neighbor connectivity $k_{nn}(k)$ is constant
  - clustering $C(k)$ is constant

Comparison with random graph models (cont.)

- skitter graph
  - most disassortative, $r = -0.24$
  - average neighbor connectivity varies by two orders of magnitude

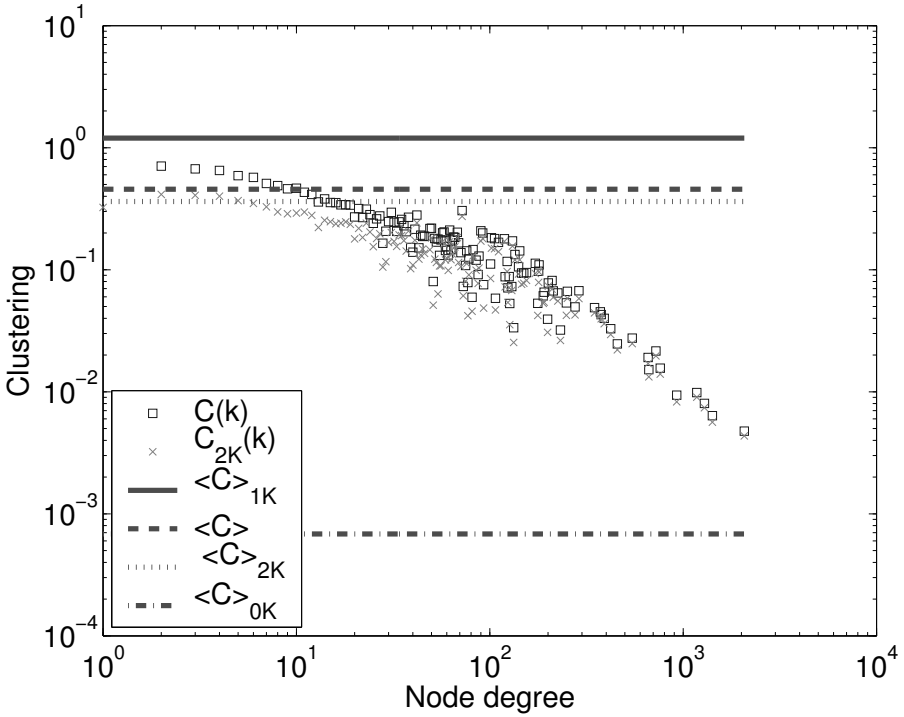
  $=>$ is not 1K-random
  cannot be approximated by PLRG

- WHOIS graph
  - almost uncorrelated, $r = -0.04$
  - average neighbor connectivity varies by a factor of two

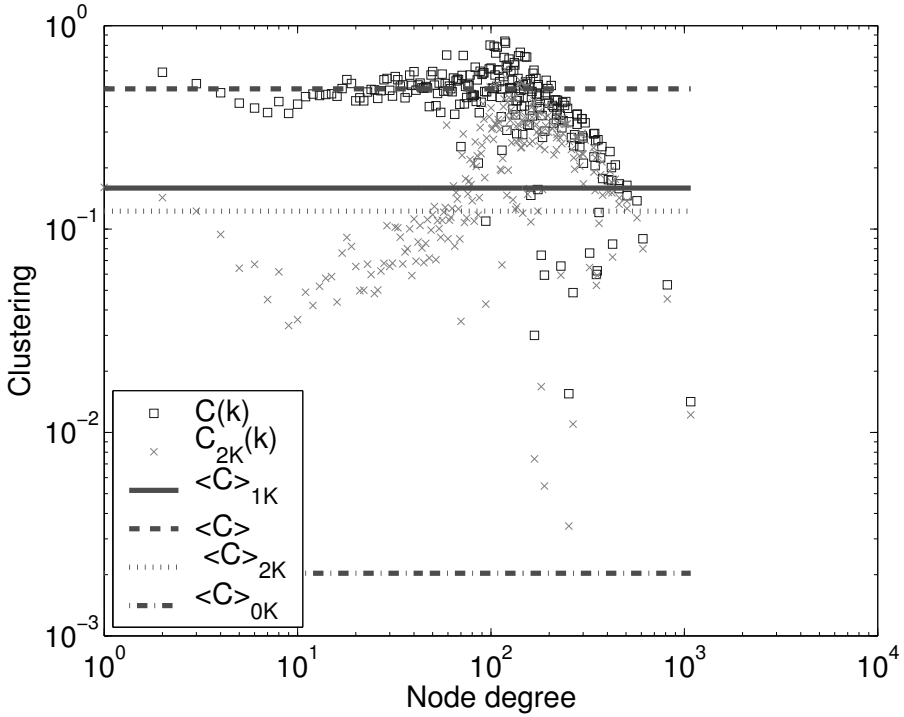
  $=>$ the closest to 1K-random
  but its degree distribution does not follow power-law

Clustering as a Measure of Model Accuracy

skitter



WHOIS



- skitter graph
  - clustering is close to 2K-random one
    $\Rightarrow$ 2K-random model reproduces skitter topology
- WHOIS graph
  - clustering is functionally different from 2K-random one
  - mean clustering is closest to 1K-random one

Conclusions

- Graphs derived from three sources of Internet topology data
  - skitter
  - BGP
  - WHOIS
- Wide range of topology metrics
- JDD $P(k_1, k_2)$ plays a definitive role
  - coarse summary statistics of JDD, $\overline{k}$ and r, explain the relative order of all other metrics
- which data source is most accurate?
  - each approximates a different view of the Internet
  - each has its own limitations and inaccuracies
  - differences are quantitative, not qualitative

$=>$ combine the reliable information from all sources for the most complete view