

Function from Structure in Communication Networks

Statistical Mechanics Day V,
June 25, 2012

Weizmann Institute, Rehovot
Scott Kirkpatrick, HUII

Inferring function from structure in social or communications networks

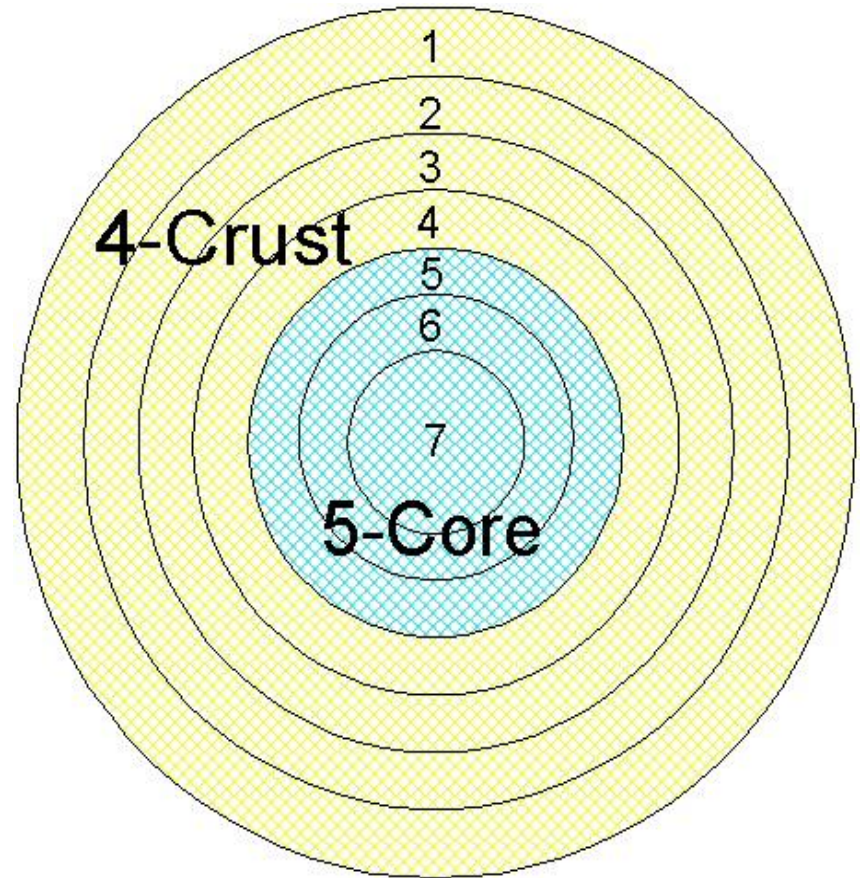
- Outline:
 - How networks are formed – social vs communications
 - Analysis tools and their complexity
 - Degree distributions, 1-pt, 2-pt
 - Betweenness and other centrality estimates
 - K-pruning
 - Observations on the Internet physical topography
 - Traditional human communication– 7 Billion telephone CDRs (all of UK, 2005)
 - Modern communication – 100M Tweets over 1.5 B links (Twitter 2010 dataset)

The Physical Internet

- Undirected links between routers with real locations
- Detected by traceroute or publication of routes
 - CAIDA, DIMES, Routeviews...
- Observed on several scales:
 - AS graph (<64K Ases), POPs, cities, IP addresses or routers
 - Only AS graph is really well understood
- Network formation governed by rules and business objectives – worldwide communications

k-Core Method

- Some definitions :
- k-Core – union of all shells with indices $\geq k$.
- k-Crust – union of all shells with indices $\leq k$.

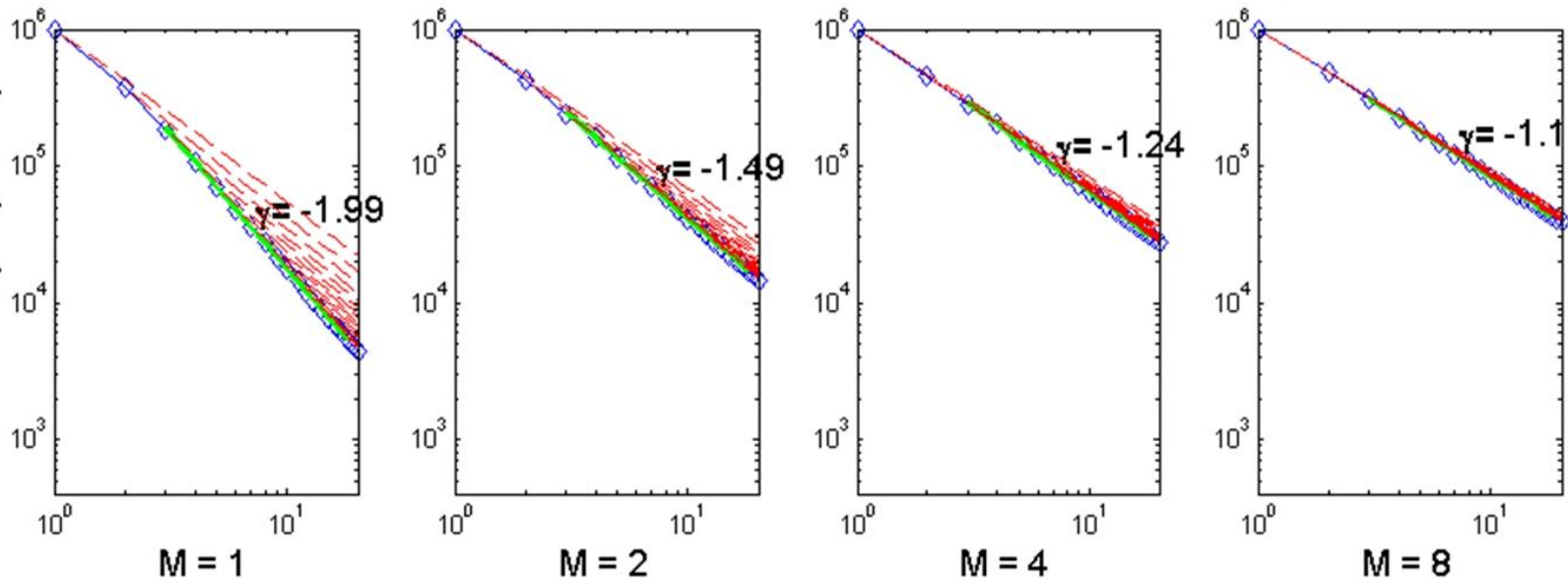


K-pruning gives a principled AS graph structure

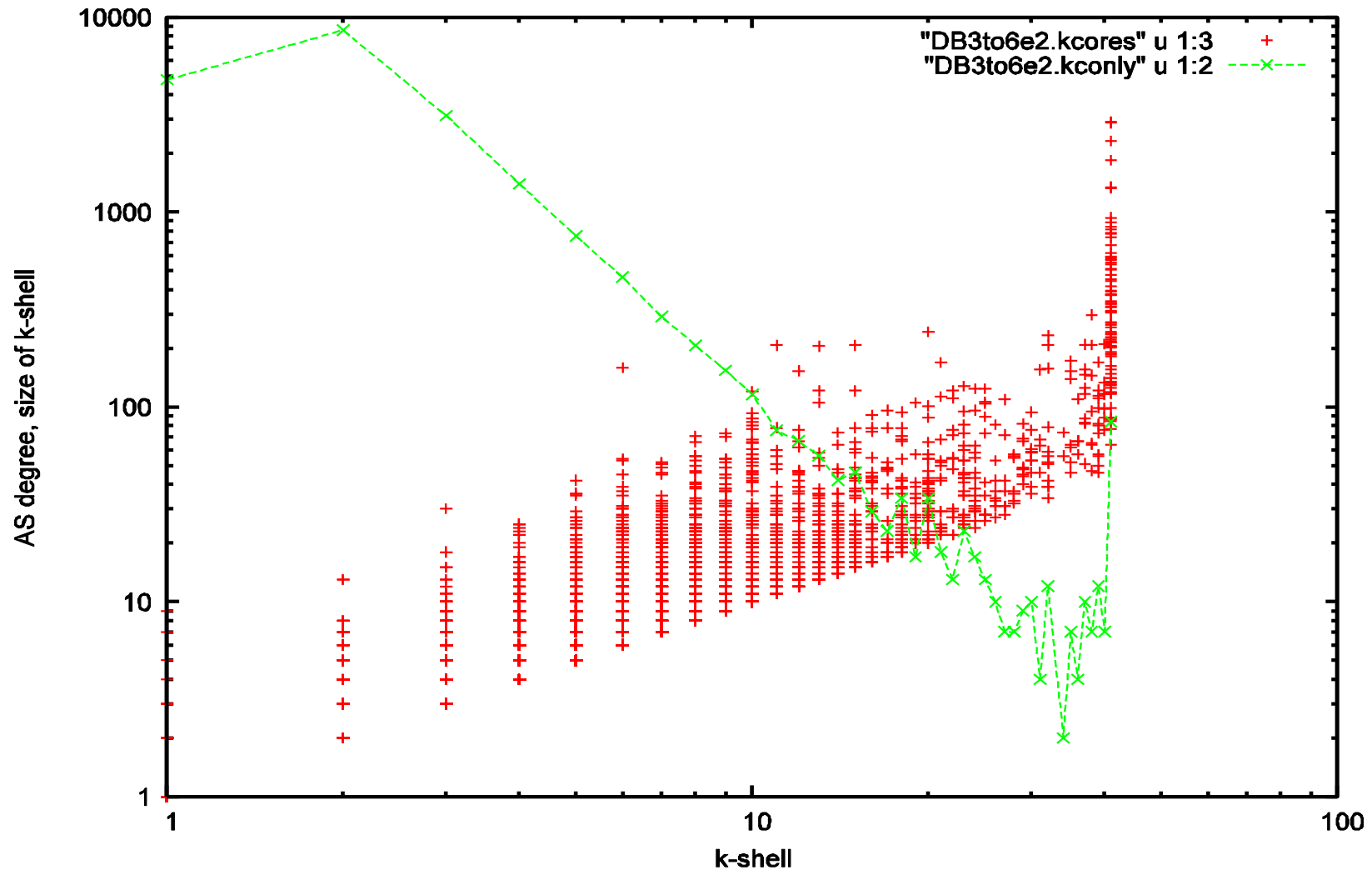
- Label graph from the edge inwards
- Order sites by their communications options
- Prune into k-shells, k-cores, k-crusts, nucleus
- Power law structure observed
 - Clearest example – preferential attachment
 - AS graph of actual internet fits this well
- Isolated “tendrils” connect only to the nucleus.

Preferential Attachment, Rome 2000

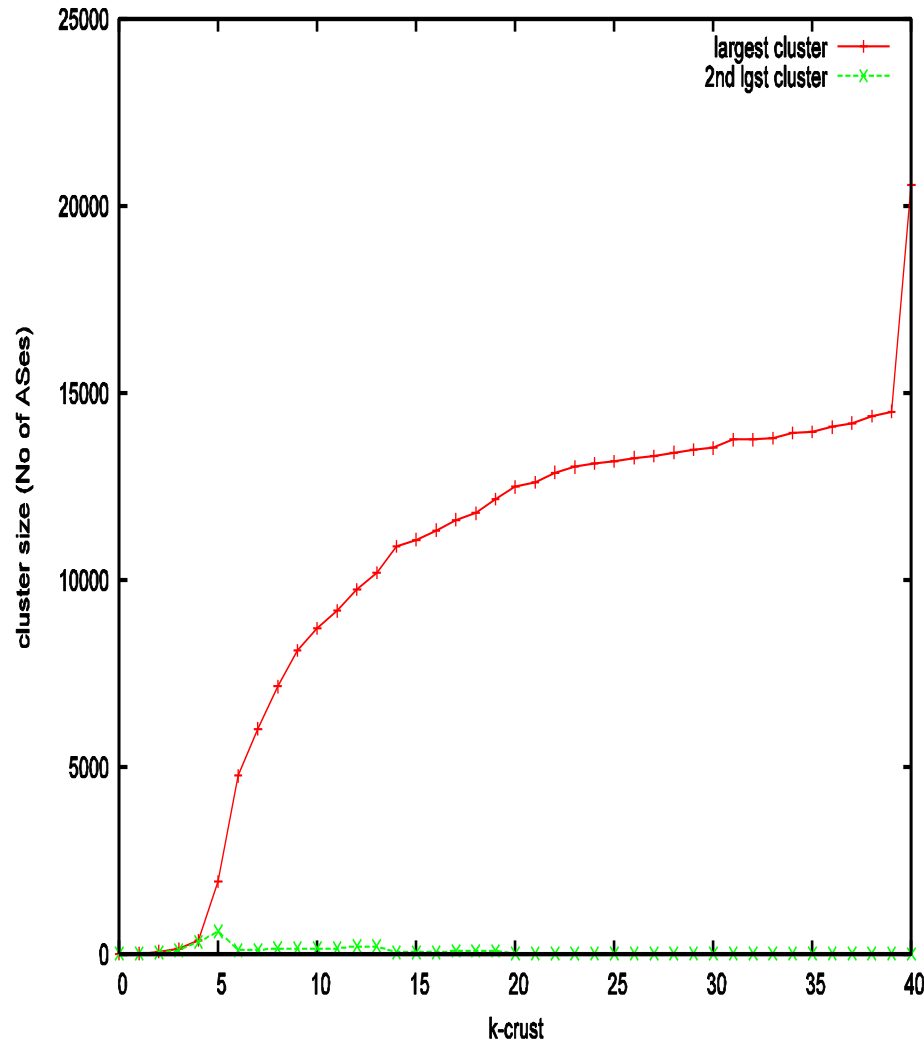
Size of the Remaining Graph , (referenced to the degree distribution)



How does original degree map into k-shell?

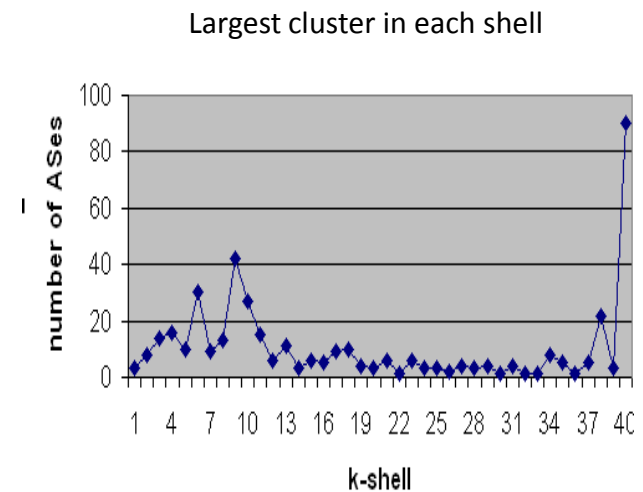


K-crusts show percolation threshold



← These are the hanging tentacles of our (Red Sea) Jellyfish

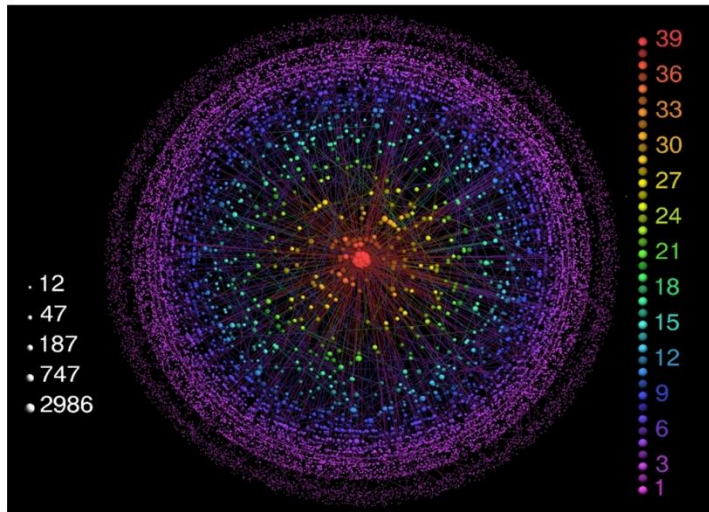
For subsequent analysis, we distinguish three components: Core, Connected, Isolated



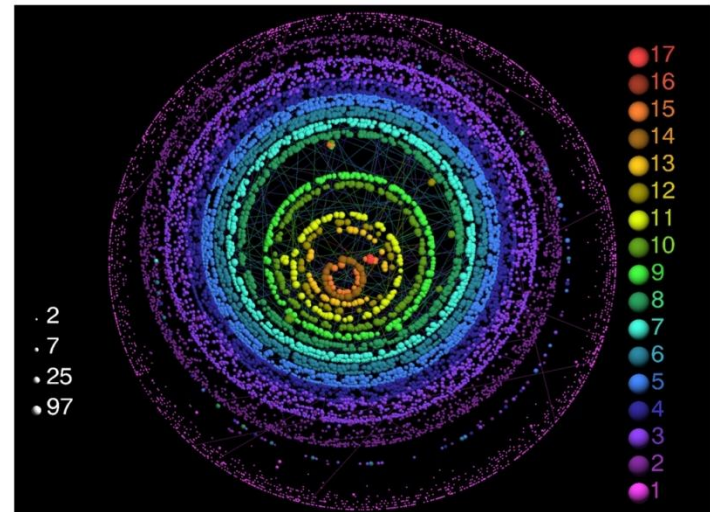
Data from 01.04.2005

Are Social Networks Like Communications Networks?

- Visual evidence that communications nets are more globally organized:
 - Indiana Univ (Vespignani group) visualization tool



AS graph, ca 2006



Movie actors' collaborations

Assortative, Disassortative, ...?

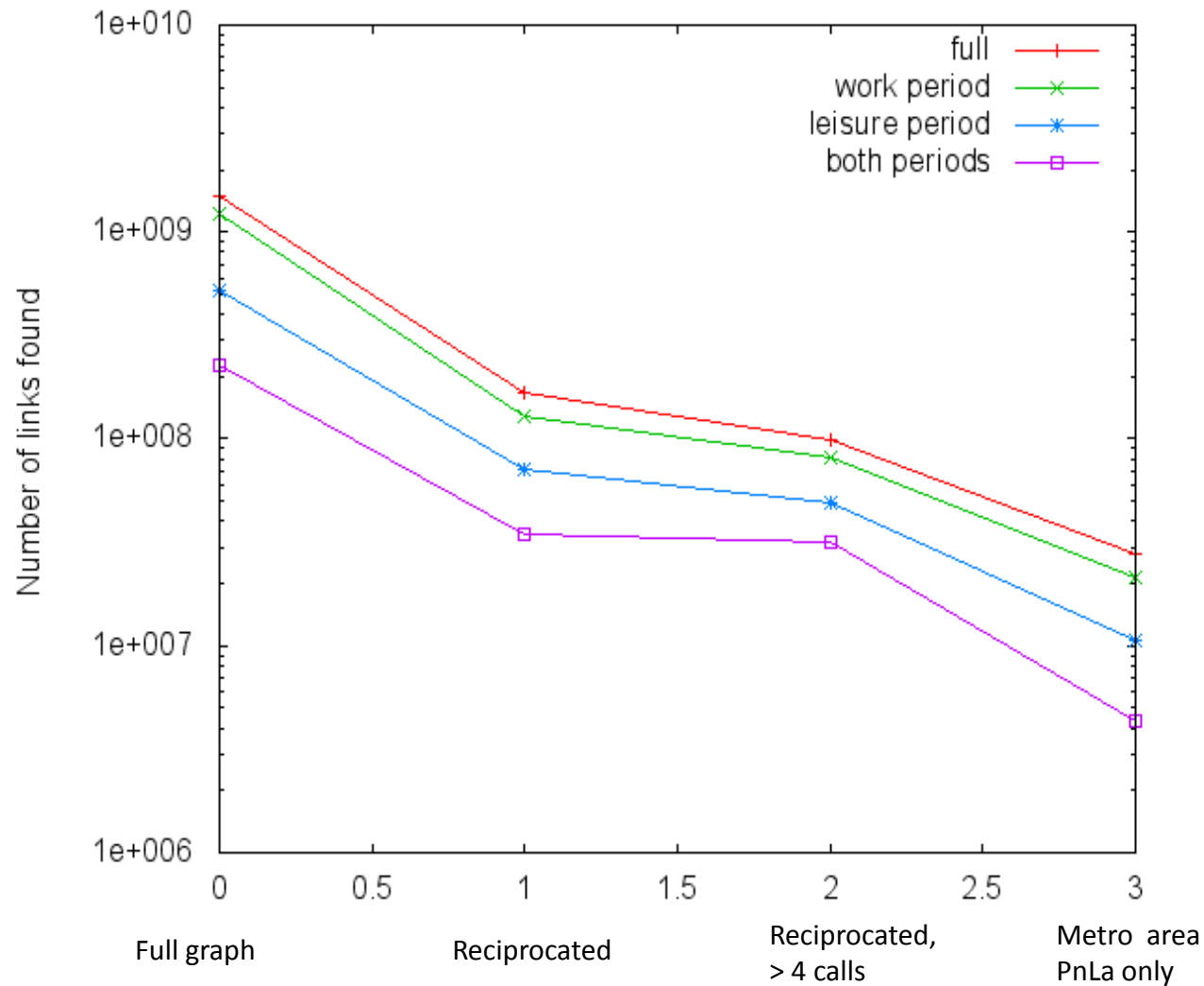
- Newman's SIAM review article distinguishes social and communications networks
 - Comm – low degree sites connect to high degree
 - Social – high degree sites cluster
- Our test case – UK CDR's for August 2005
 - Social, communications, or a bit of both?
 - If both, do we see the average behavior or can we separate disassortative parts of the net from other?

Telephone call graphs (“CDRs”)

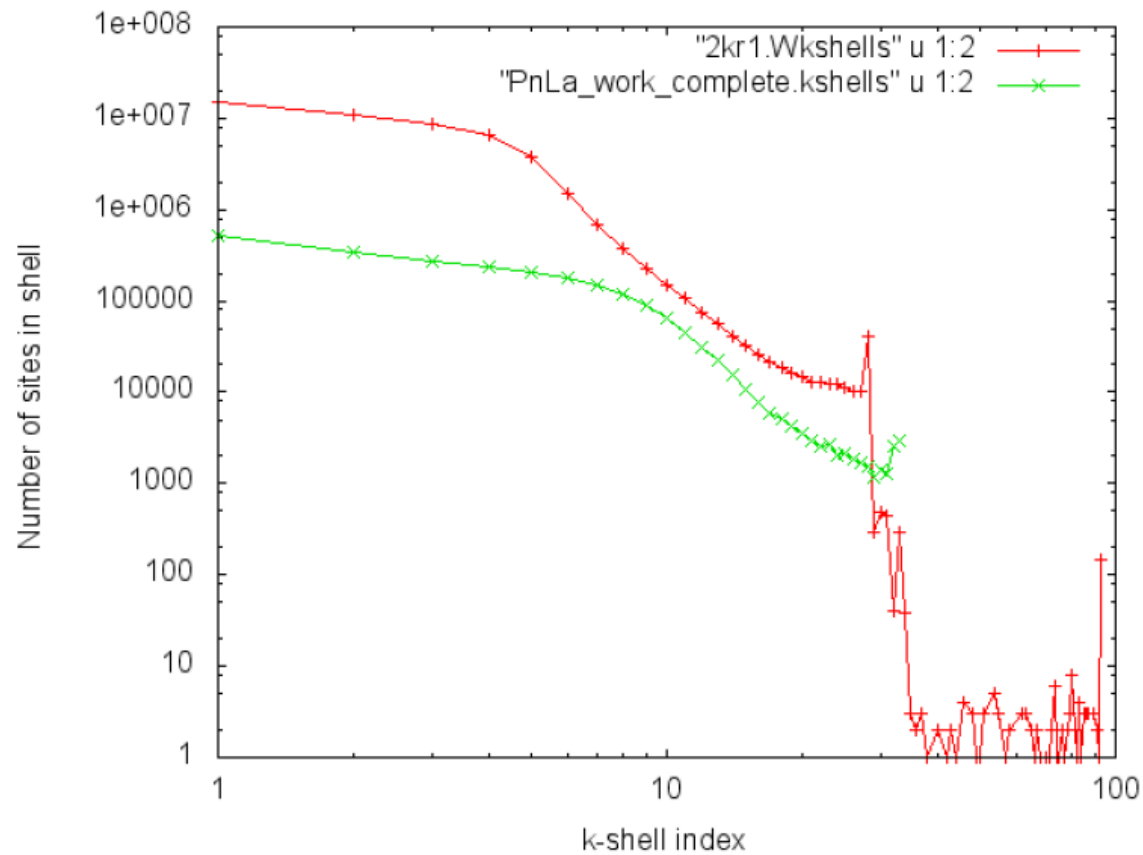
Can be studied on several scales

7 B calls, over
28 days, Aug
2005

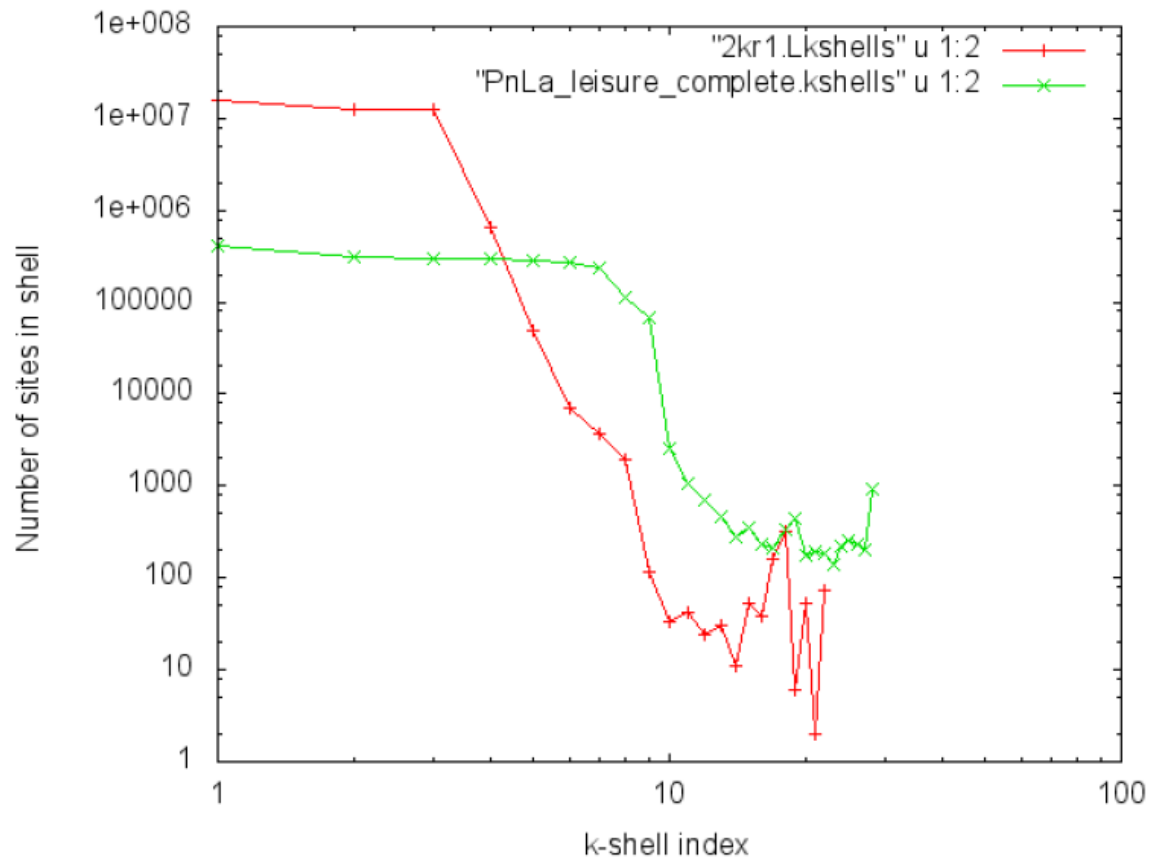
Cebrian,
Pentland,
SK



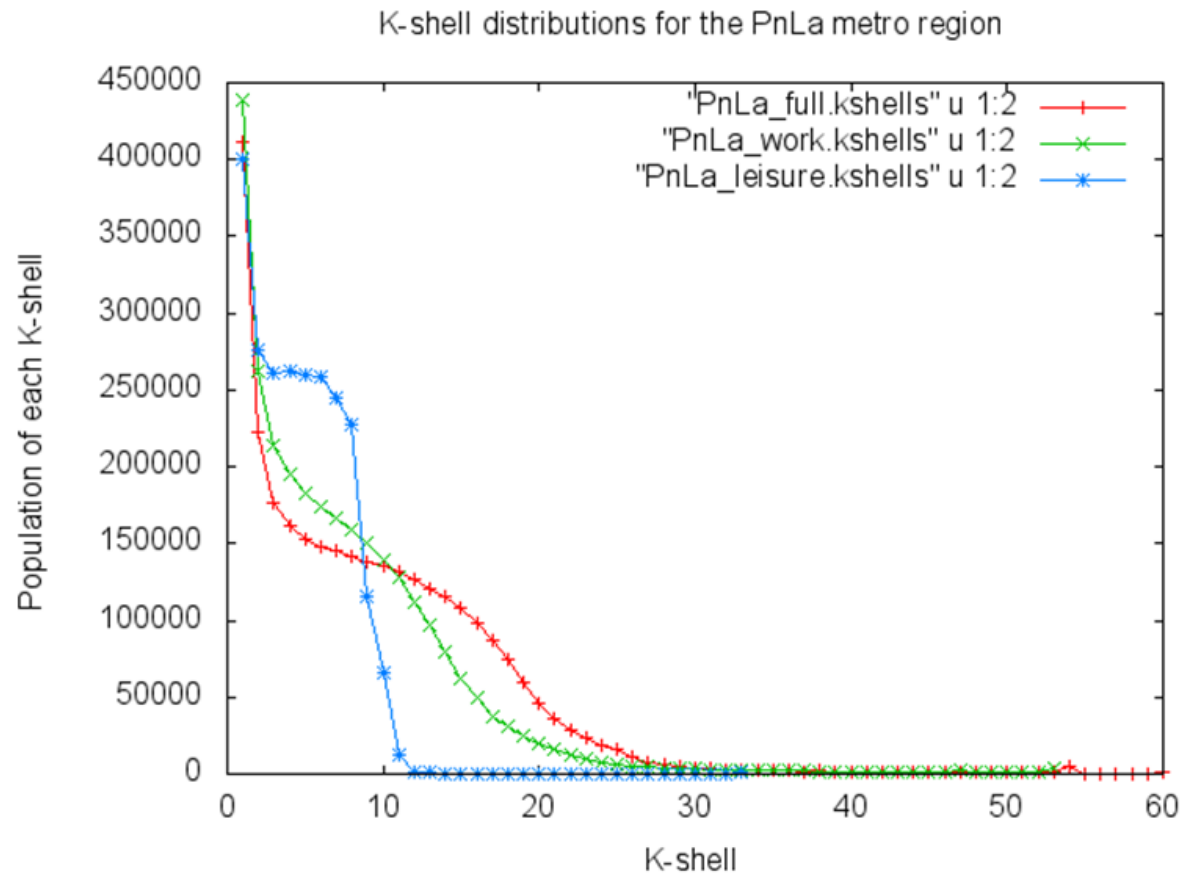
Work k-shells in CDR network



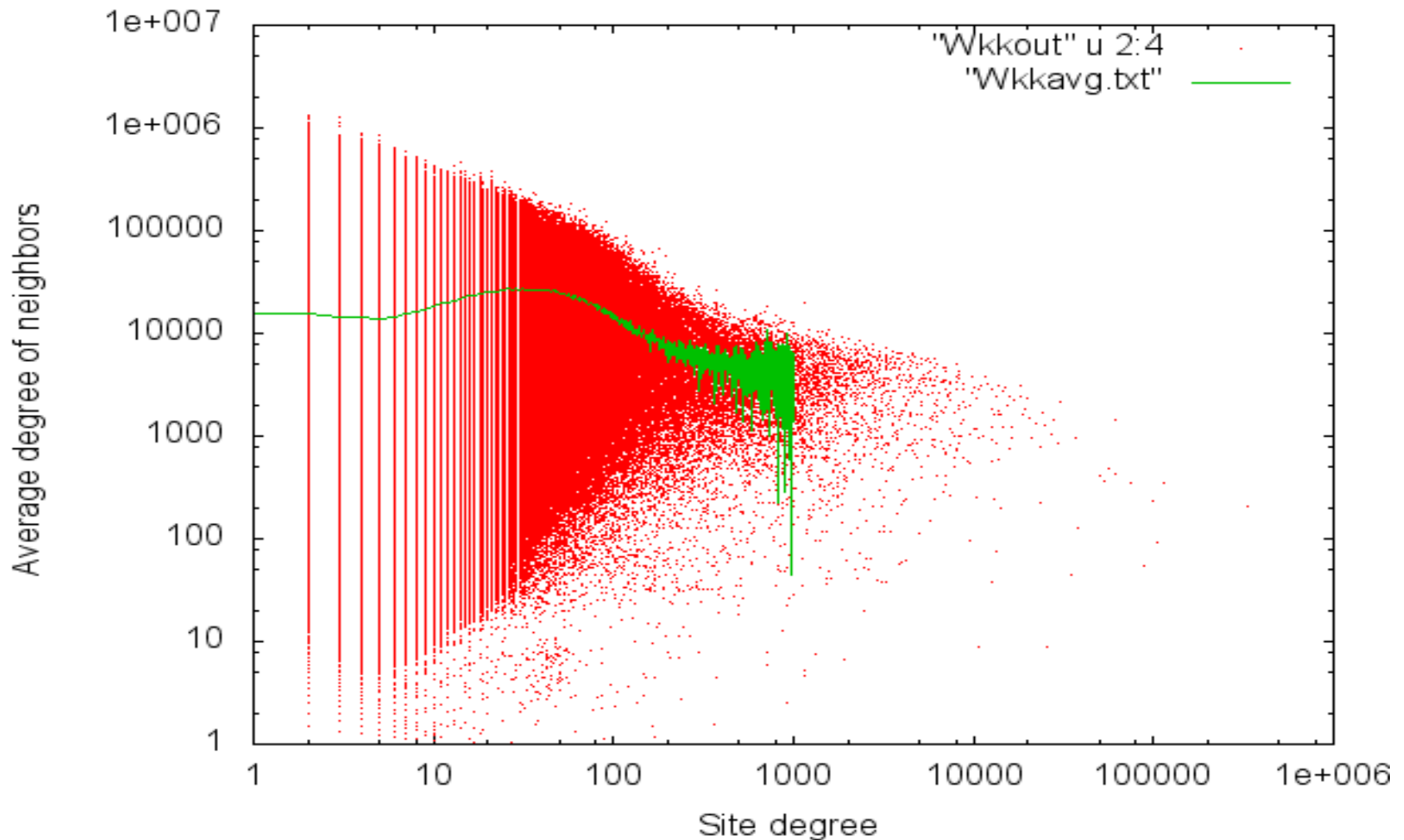
Leisure k-shells in CDR network



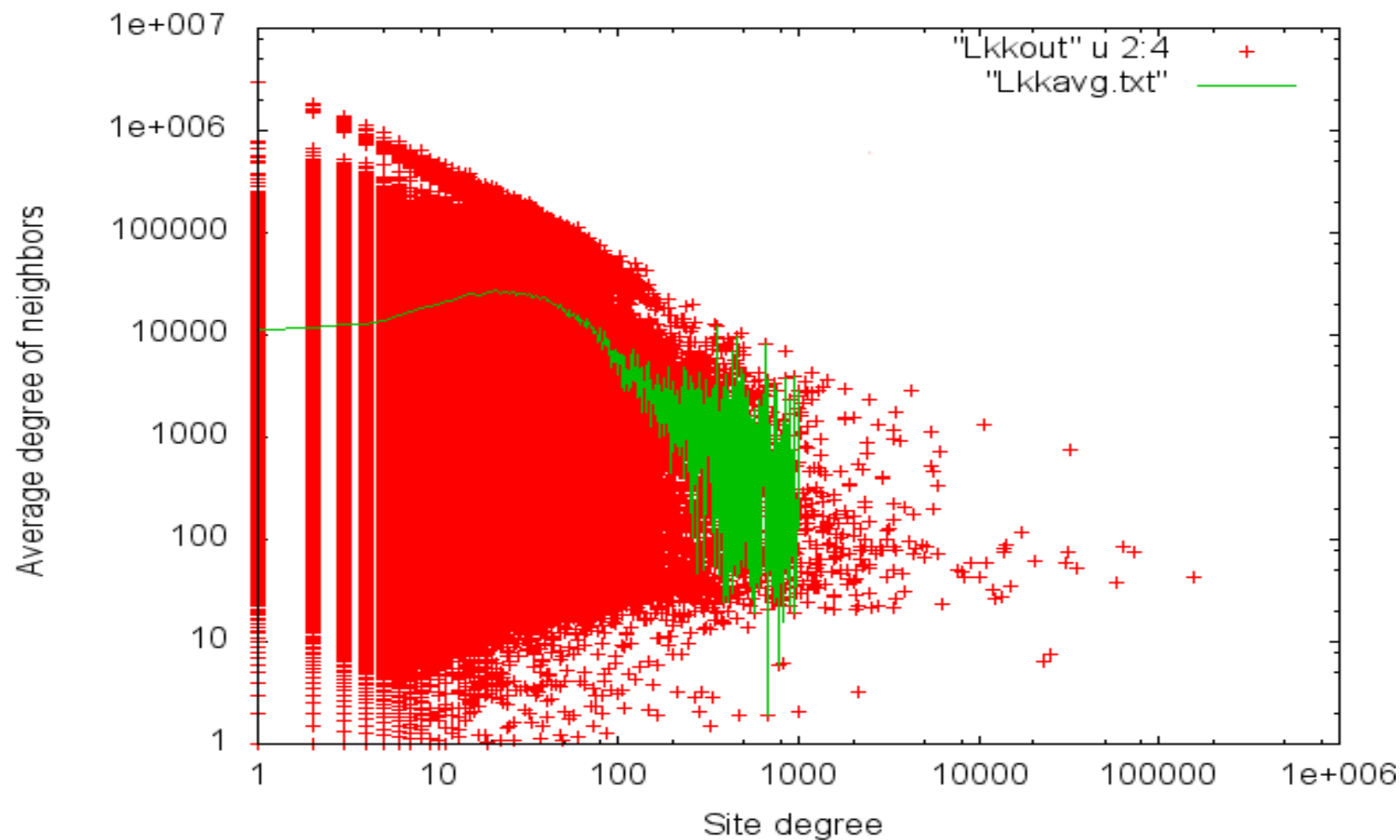
“PnLa” metro region (linear scale)



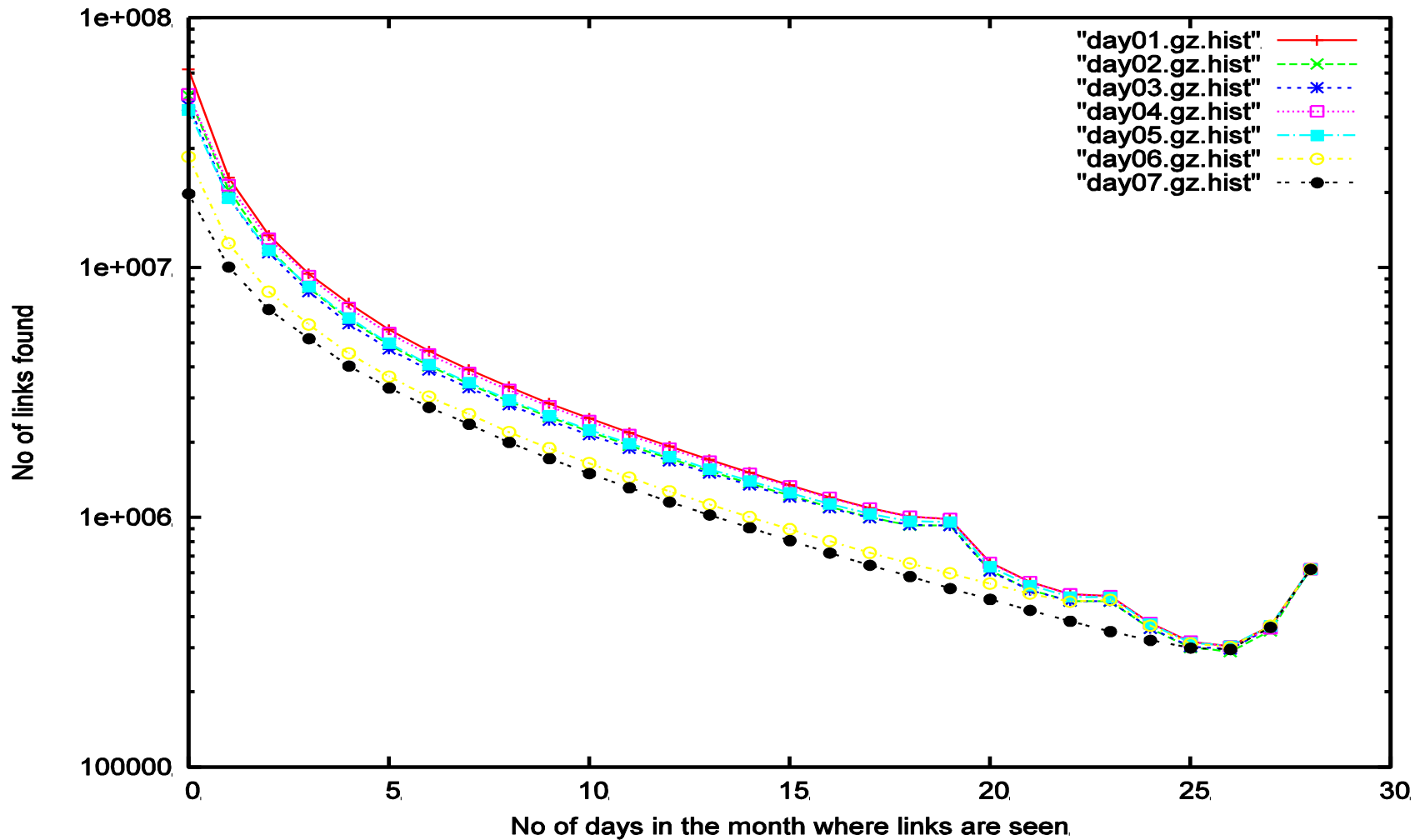
Work full network degree-degree correlations – a mixed picture



Leisure also mixed



Evanescent and Persistent Links



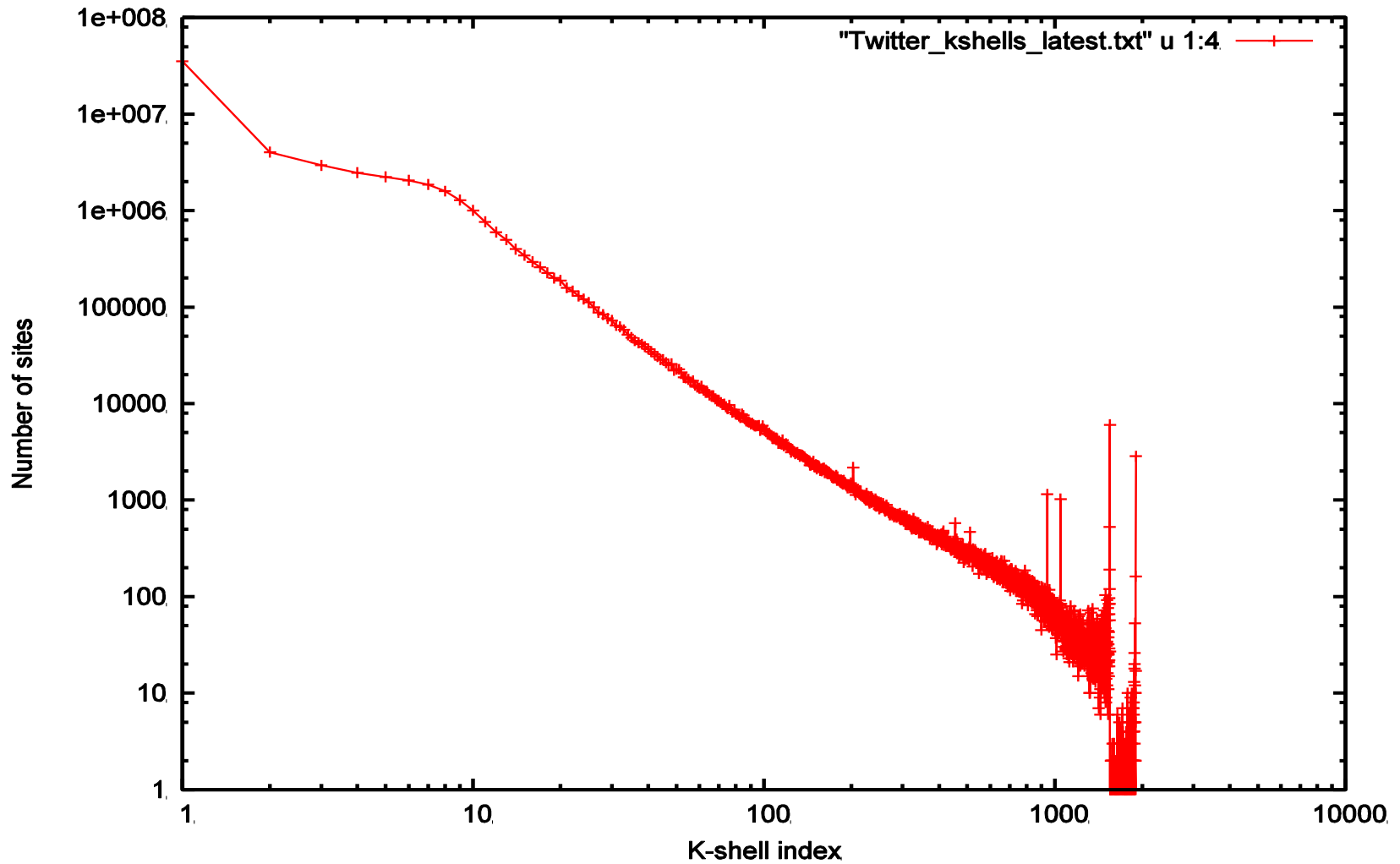
Conclusions (tentative)

- CDR network has elements of both social and communications typical structure
- Low k sites are social, while high k sites (some of them) have communications purpose
- Other high k sites are ubiquitous
- Nature of the last core uncertain
- Is there a “big brother” among us?
- Need efficient (sampling) tools for analysis

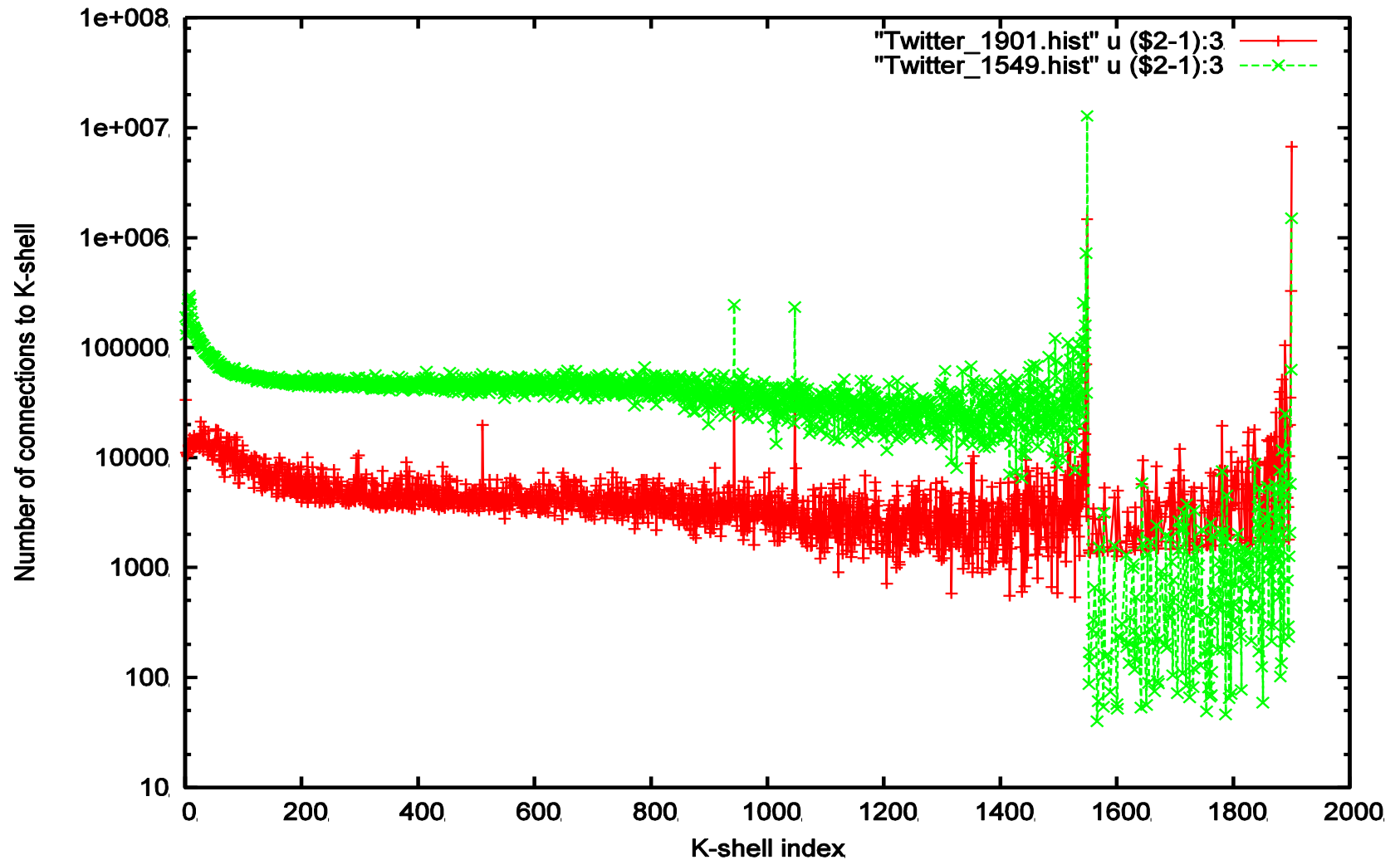
Next steps, next datasets

- CDR phone call records (120 M users 6-7B calls)
- Twitter graph (40 M nodes, 1.4 B edges)
- Grounds for exploring the temporal structure of communications between people
 - Today vs 5+ years ago, has twitter changed the nature of communications
 - Burst behavior and control structures within each dataset
- Requires HPCC tools and skills: Hadoop, Graphlab...

The Twitter K-shell structure



Twitter network's two nuclei

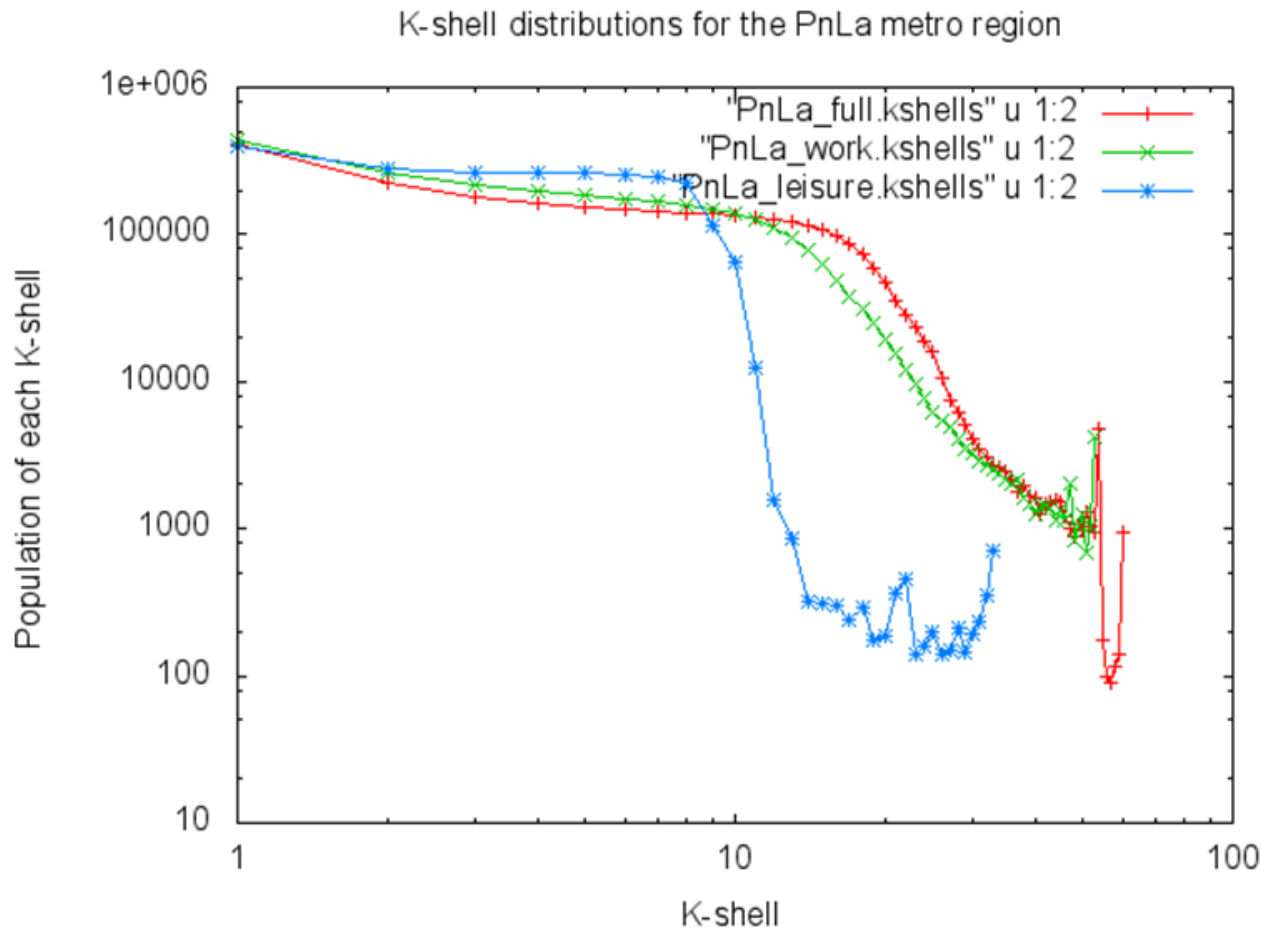


Acknowledgements

- Sorin Solomon, Avishalom Shalitt
- Yuval Shavitt, Eran Shir, Udi Weinsberg
- Shai Carmi, Shlomo Havlin
- Manuel Cebrian, Sandy Pentland, Alex Kulakovsky, Danny Bickson

Charts removed for brevity

PnLa metro region (loglog scale)



3B links, 1.2B IDs require simplified betweenness

- Subsampling required
- Traffic, from edge to edge sites (select 1000)
- Choose a “nucleus” set to study
- Computation:
 - Find distances from each nucleus site to all edge sites $d(n,i)$
 - Find all edge to edge distances $d(i,j)$
 - Score one point for site n if $d(i,j) = d(n,i) + d(n,j)$

Edge to edge traffic

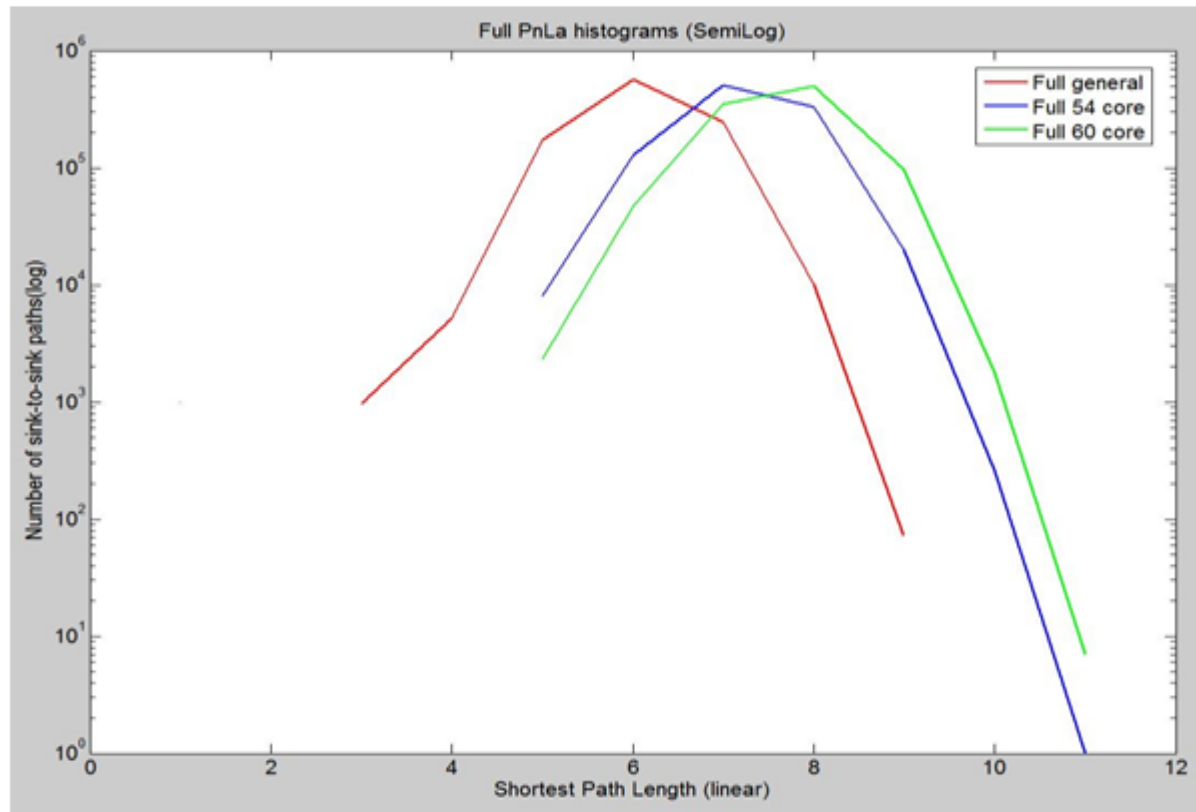


Fig. 2a Edge-to-edge shortest path length distributions for full PnLa network.

Leisure traffic shows same pattern

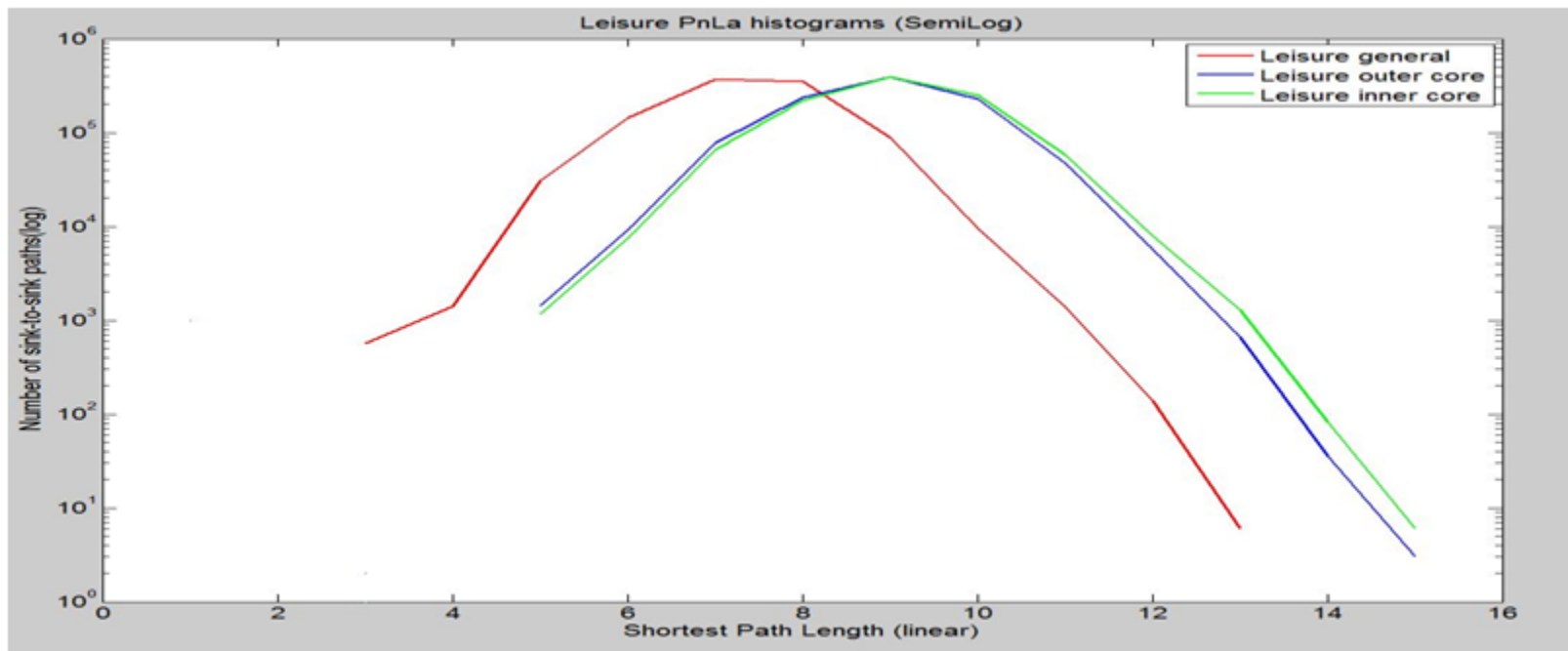
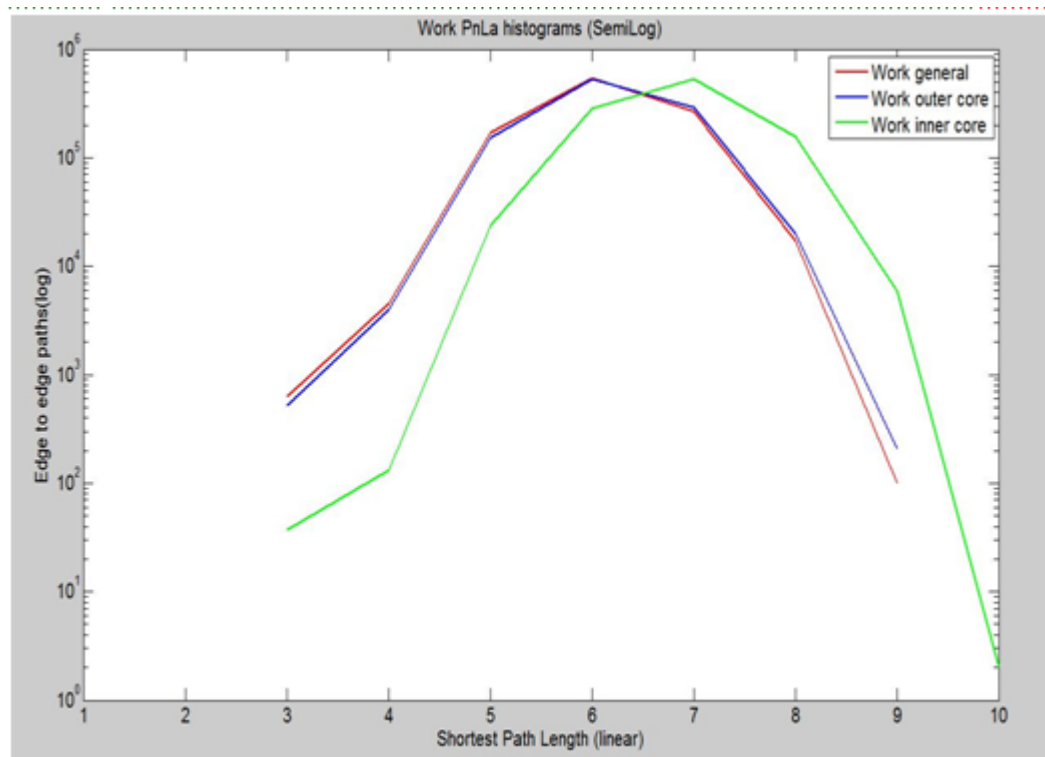
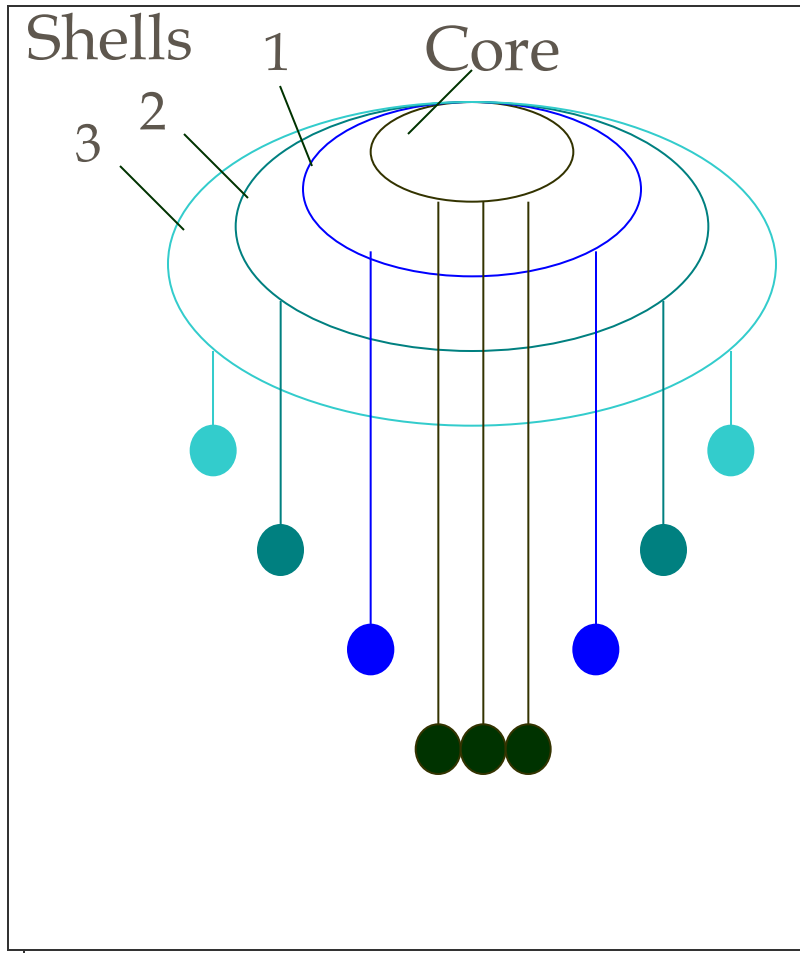


Fig 2b. Edge-to-edge shortest path length distribution for Leisure PnLa network.

Work traffic uses the outer nucleus

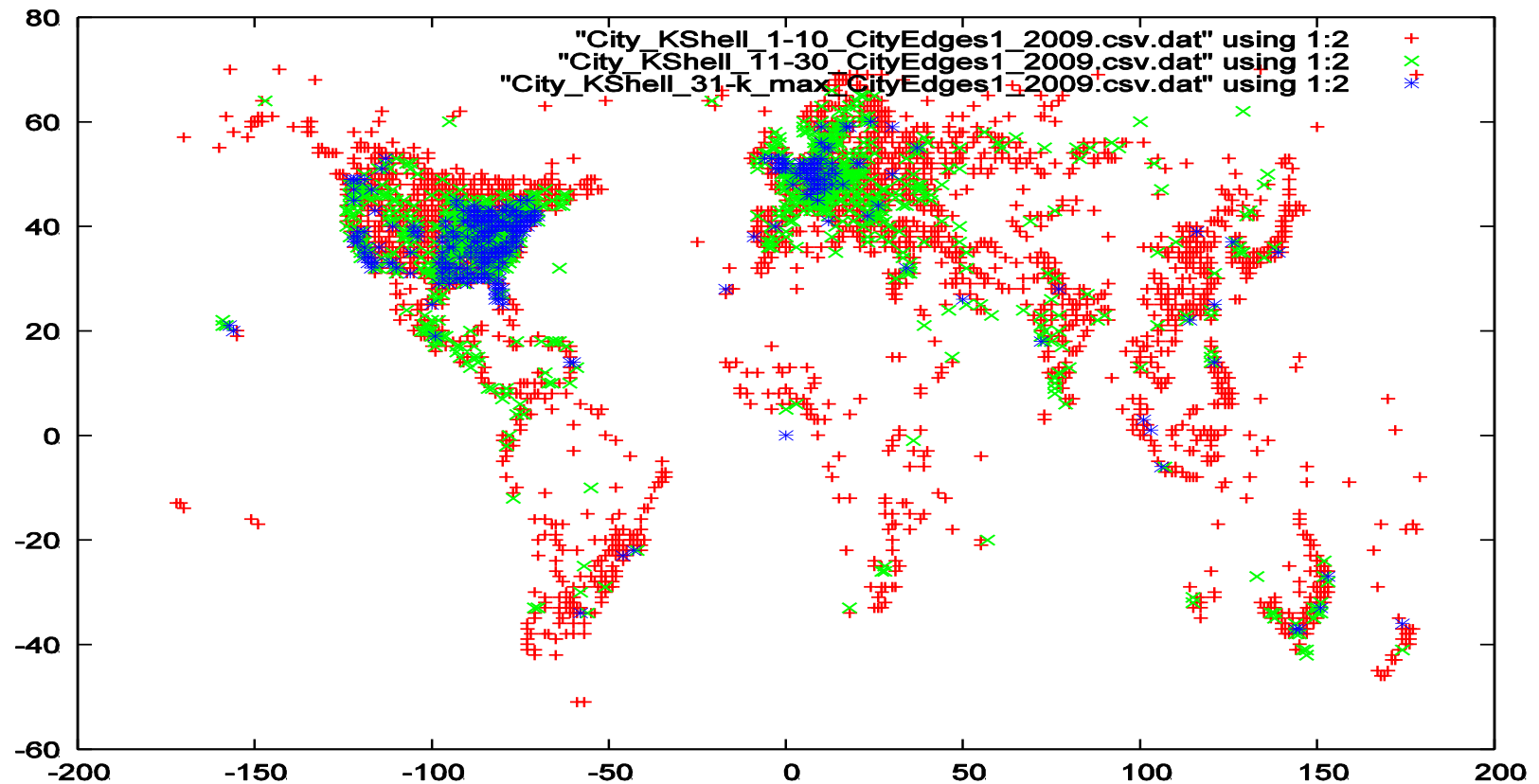


Michailis Faloutsos' Jellyfish

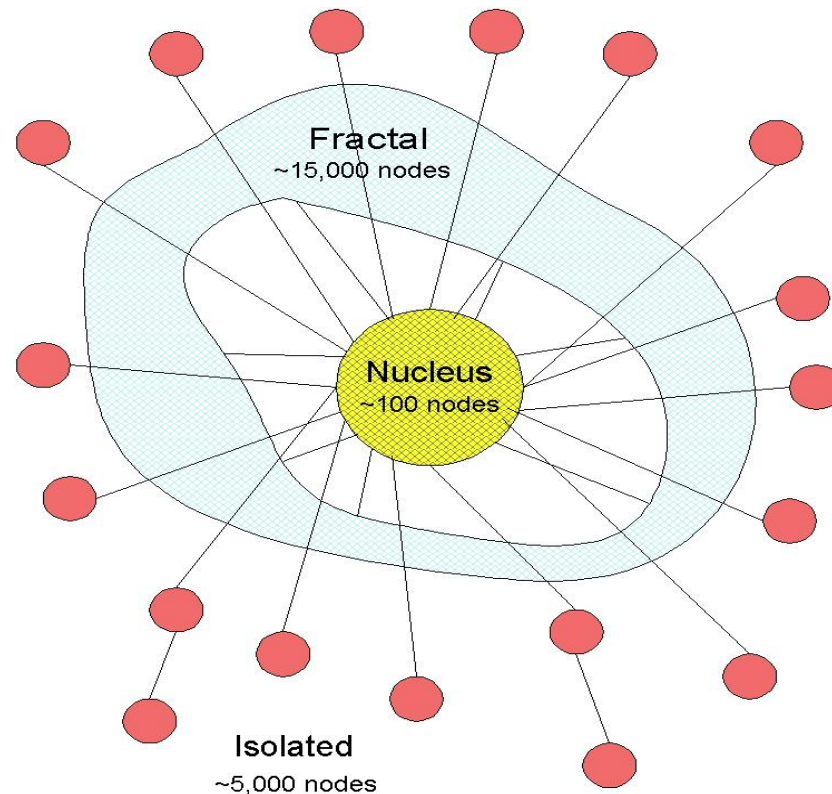


- Highly connected nodes form the core (“Tier One nets”)
- Each Shell: adjacent nodes of previous shell, except 1-degree nodes
- **Importance** decreases as we move away from core
- 1-degree nodes hanging
- No principled way of defining the core of the jellyfish (max-clique or extra-dense subset?)

City locations permit mapping the physical internet



Meduza (מדוזה) model



This picture has been stable from January 2006 ($k_{\max} = 30$) to present day, with little change in the nucleus composition. The precise definition of the tendrils: those sites isolated from the largest cluster in all the crusts – they connect only to the core.