dK-series and hidden hyperbolic metric spaces

Dmitri Krioukov

dima@caida.org

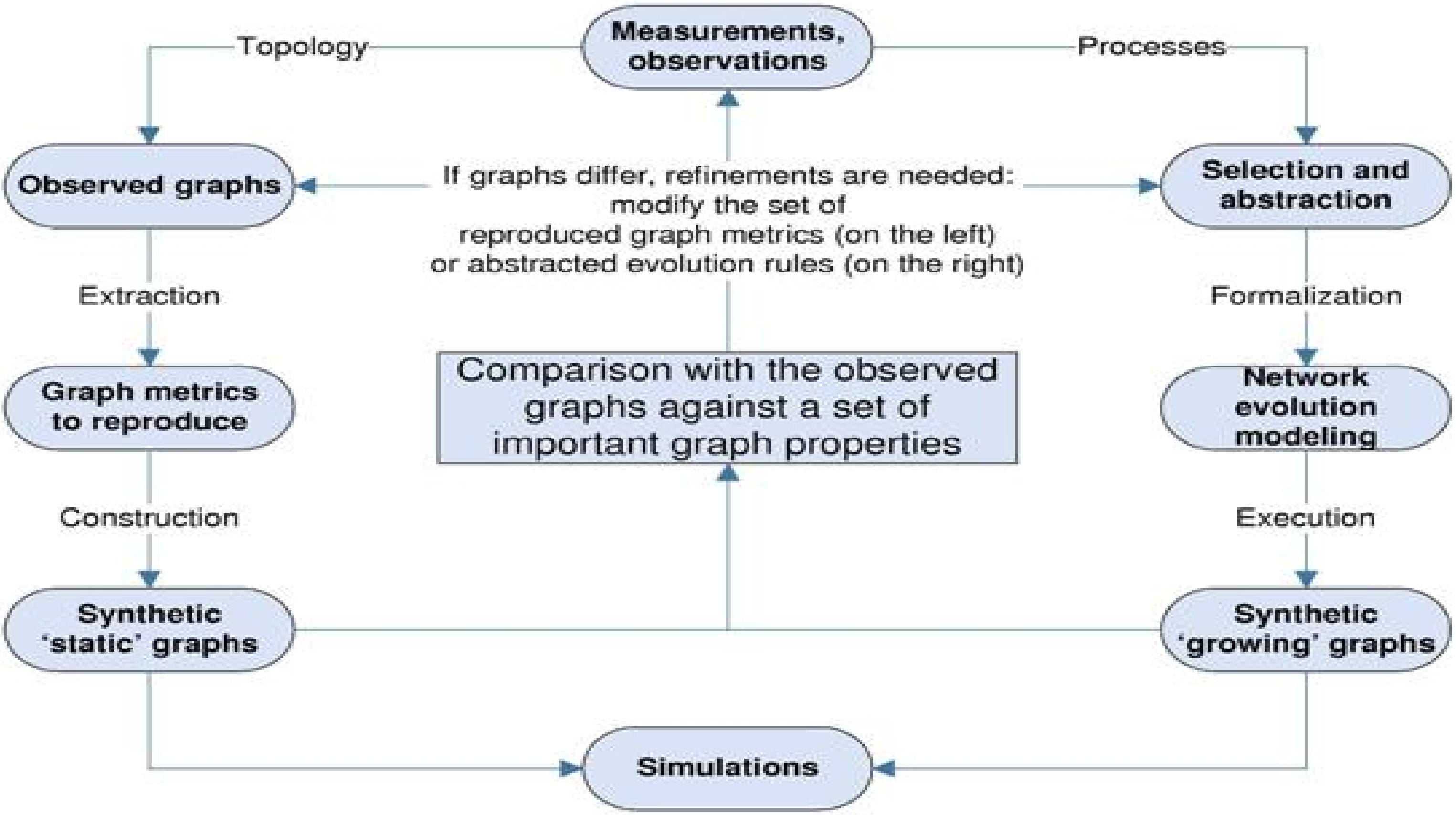
A. Jamakovic, P. Mahadevan, F. Papadopoulos, K. Fall, A. Vahdat, and M. Boguña

Telefonica, June 26th, 2009

Motivation: topology analysis and generation

- New *routing* and other protocol design, development, testing, etc.
  - Analysis: performance of a routing algorithm strongly depends on topology, the recent progress in routing theory has become topology analysis
  - Generation: empirical estimation of scalability: new routing might offer X-time smaller routing tables for today but scale Y-time worse, with $Y >> X$
- Network robustness, resilience under attack, worm spreading, etc.
- Traffic engineering, capacity planning, network management, etc.
- Motifs: are they really functional building blocks?
- In general: local vs. global network properties, network structure vs. function, and “what if” scenarios, better predictive power

Network topology research



Important topology metrics

- Spectrum
- Distance distribution
- Betweenness distribution
- Community structure
- Motif distribution
- Degree distribution
- Assortativity
- Clustering

Problems

- No way to reproduce most of the important metrics simultaneously
- No guarantee there will not be any other/new metric found important

Our approach

- Look at inter-dependencies among topology characteristics
- See if by reproducing most basic, simple, but not necessarily practically relevant characteristics, we can also reproduce (capture) all other characteristics, including practically important
- Try to find the one(s) defining *all others*

Outline

- Introduction
- *dK*-*:
  - *dK*-distributions
  - *dK*-series
  - *dK*-graphs
  - *dK*-randomness
  - *dK*-generator (Orbis)
- *dK*-randomness of real networks
- Hidden hyperbolic metric spaces as an explanation
- Conclusion

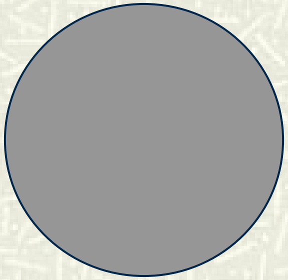
The main observation ☺

Graphs are structures of *connections* between nodes

dK-distributions as a series of graphs’ *connectivity* characteristics

0K

Average degree $<k>$

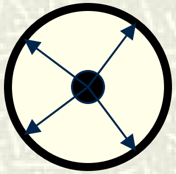


1K

Degree distribution *P(k)*

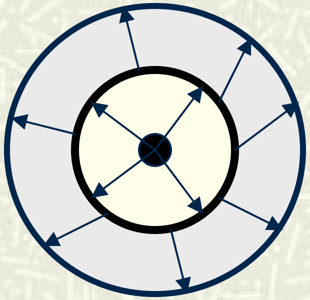
2K

Joint degree distribution $P(k_1,k_2)$

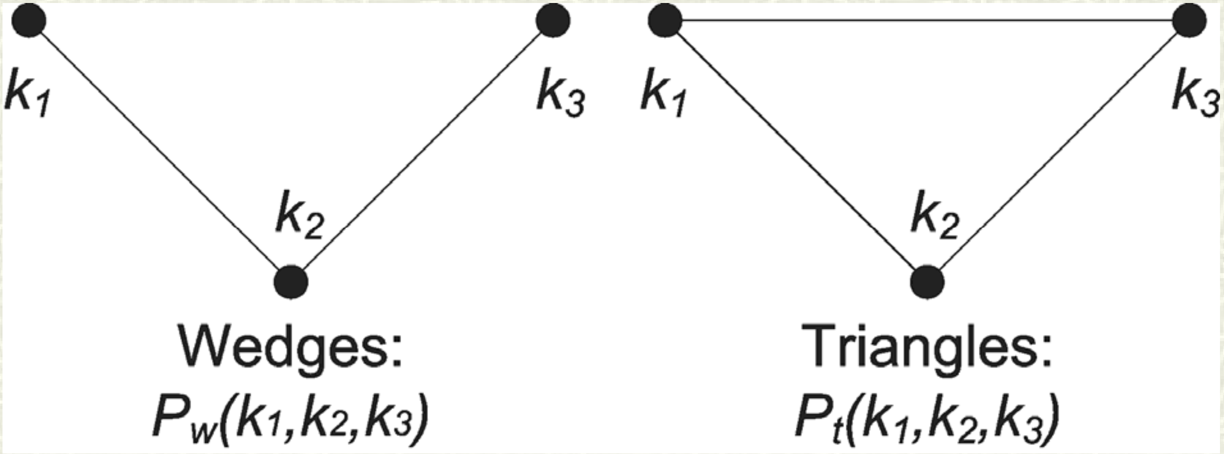


3K

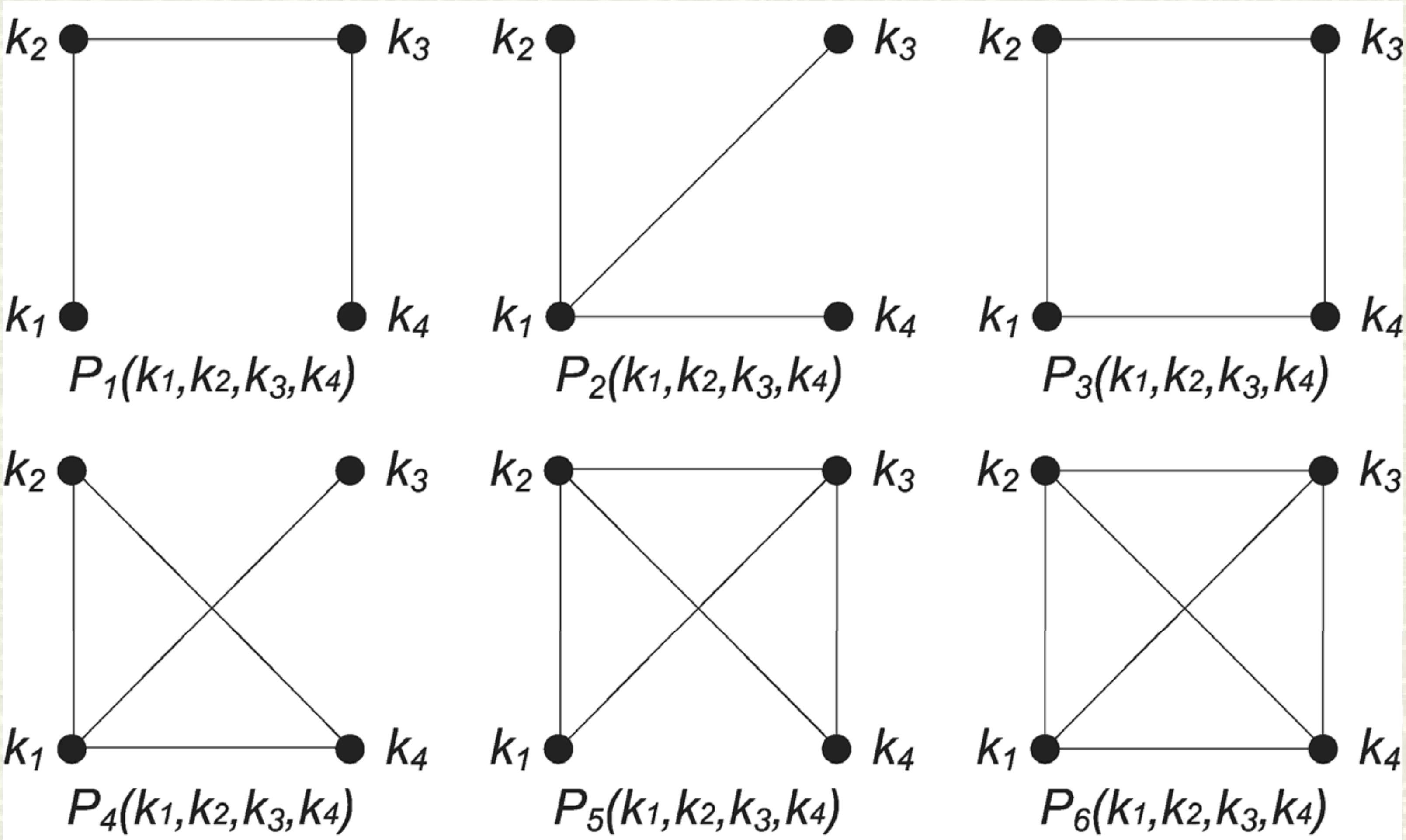
“Joint edge degree” distribution $P(k_1,k_2,k_3)$



3K, more exactly



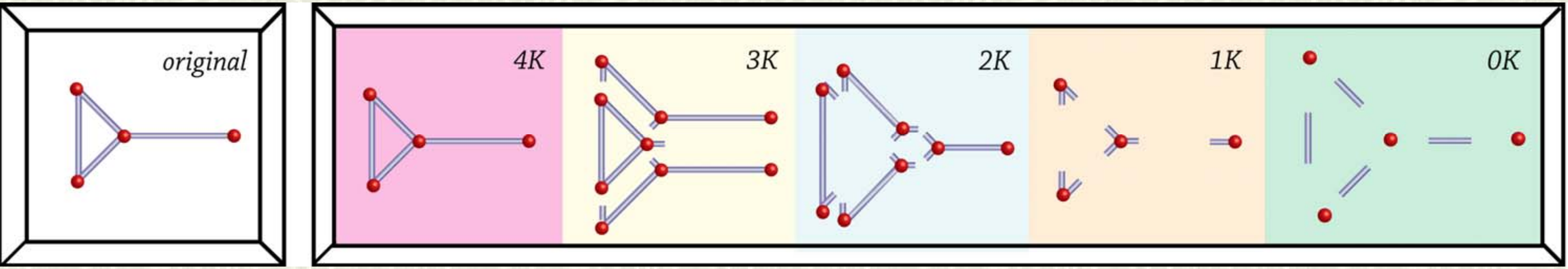
4K


k2
k3
k1
k4
P1(k1,k2,k3,k4)
P2(k1,k2,k3,k4)
P3(k1,k2,k3,k4)
P4(k1,k2,k3,k4)
P5(k1,k2,k3,k4)
P6(k1,k2,k3,k4)

Definition of *dK*-distributions

dK-distributions are degree correlations within simple connected graphs of size *d*

dK-decomposition example



4K	3K	2K	1K	0K
$N_4(3,2,1,2)=1$	$N_t(2,2,3)=1$	$N(2,2)=1$	$N(1)=1$	$k=2$
	$N_w(2,3,1)=2$	$N(2,3)=2$	$N(2)=2$	
		$N(1,3)=1$	$N(3)=1$	

Definition of *dK*-series P_d

Given some graph *G*, graph *G'* is said to have *property* P_d if *G'*'s *dK*-distribution is the same as *G*'s

Definition of *dK*-graphs

dK-graphs are graphs having property P_d

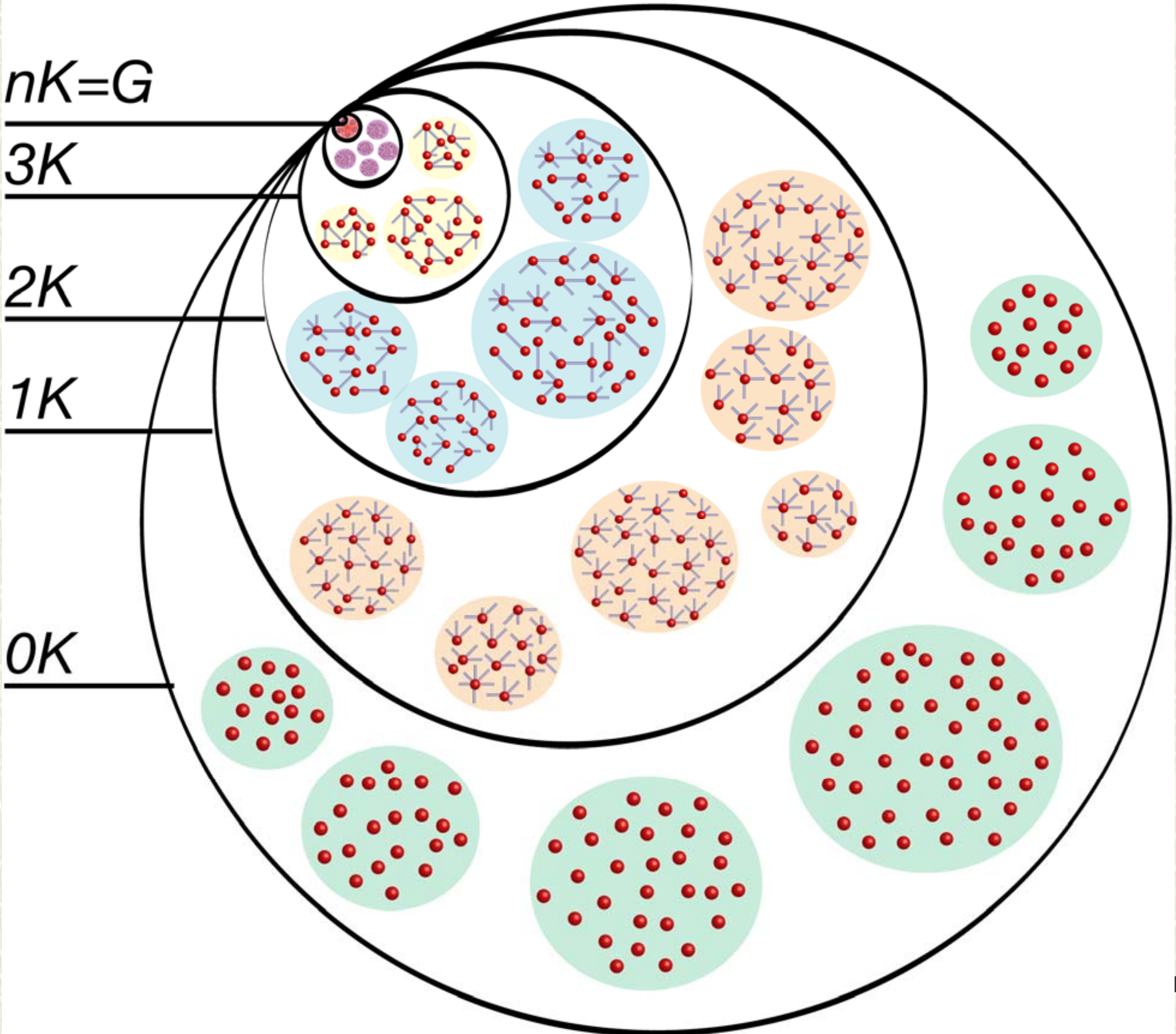
Nice properties of properties P_d

- **Inclusiveness**: if a graph has property P_d, then it also has all properties P_i, with $i < d$ (*dK*-graphs are also *iK*-graphs)
- **Convergence**: the set of graphs having property P_n consists of only one element, *G* itself (*dK*-graphs converge to *G*)
- **Constructability**: we can construct graphs having properties P_d (*dK*-graphs)

Convergence…

…guarantees that *all* (even not yet defined!) graph metrics can be captured by sufficiently high *d*

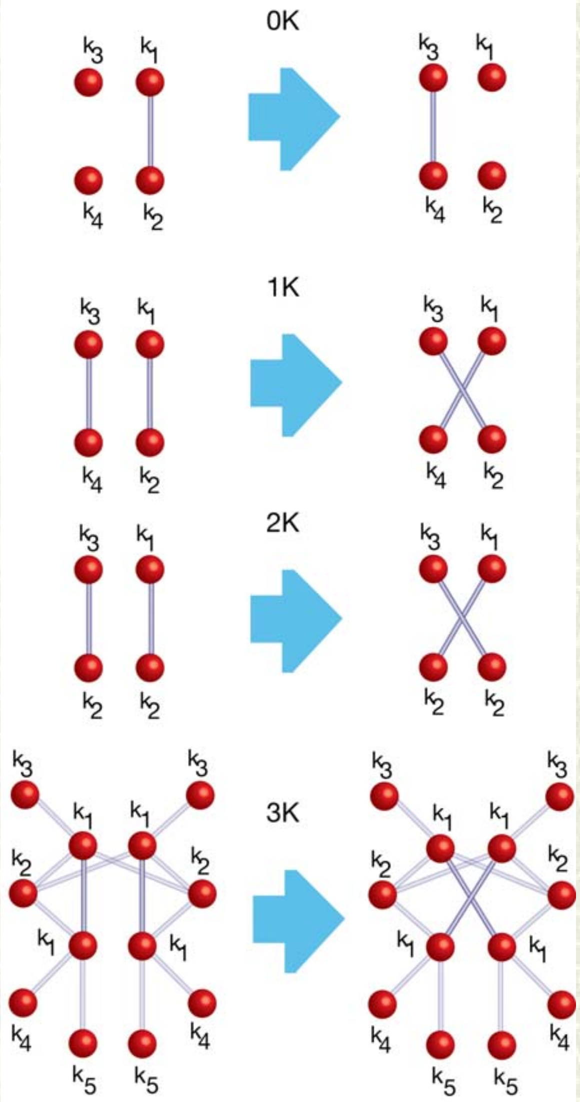
Inclusiveness and convergence



dK-random graphs vs. *dK*-graphs

- *dK*-graph is a graph that has the same *dK*-distribution as a given graph *G* (*strict definition*)
- *dK*-random graph is a "maximally random" *dK*-graph (*non-strict definition*, but very useful in practice)
  - *dK*-random graph is a graph that has the same *dK*-distribution as *G* but that is random in other respects
  - constructing *dK*-graphs, we usually construct *dK*-random graphs
  - to construct *dK*-non-random graphs, we have to inventively modify the construction procedures…

dK-randomization: random rewiring preserving the *dK*-distribution



- *dK*-randomizing a given graph *G*, we obtain its *dK*-random counterparts
- These *dK*-random graphs are always similar to each other
- Graph *G* itself is called *dK*-random if it's similar to its *dK*-random counterparts

dK-generator (Orbis)

- Establish how *dK*-random a given network *G* is, i.e., find the minimum *d* s.t. *G* is *dK*-random
- Given a *dK*-distribution (*G* no longer needed!), construct *dK*-random graphs:
  1. extract the *1K*-distribution from the *dK*-distribution
  2. construct a *1K*-random graph (many methods exist)
  3. done if *d=1*, or set *i=2* otherwise
  4. extract the *iK*-distribution from the *dK*-distribution
  5. perform *(i-1)K*-preserving iK-*targeting* rewiring, accepting each rewiring step if it moves the graph's *iK*-distribution closer to the target extracted *iK*-distribution
  6. done if *i=d*, or set *i=i+1* otherwise and go to step 4

Problem

- Complexity of *dK*-series grows hyper-exponentially with *d* – the dominating contribution is from the number of non-isomorphic graphs of size *d*
- So, how *dK*-random are real networks???

Outline

- Introduction
- *dK*-*
- *dK*-randomness of real networks
  - Networks considered
  - Methodology
  - Internet
  - Web of trust
- Hidden hyperbolic metric spaces as an explanation
- Conclusion

Networks considered

- Communication: the Internet
  - AS-level (skitter)
  - “Router”-level (HOT)
- Social:
  - Web of trust (PGP)
  - Paper co-authorship network (arXiv)
- Biological:
  - Protein interactions (yeast *Saccharomyces cerevisiae*)
- Transportation:
  - US airport network
- Technological:
  - Western US power grid
- Few others
  - including a dolphin acquaintance network!

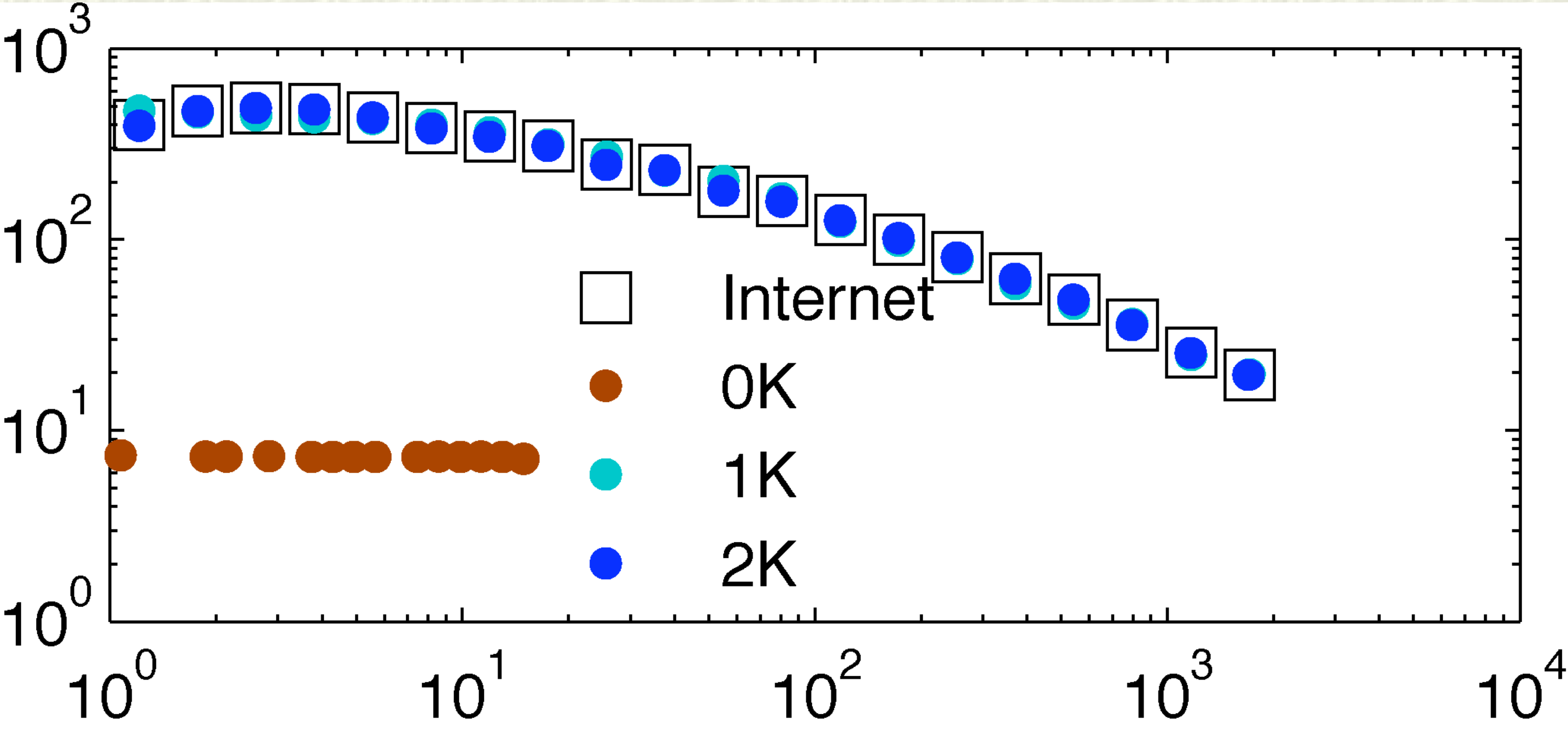
Main finding

- All networks are *3K*-random at most
  - AS-level Internet is *1K*-random
  - Airport network is *2K*-random
- Except the power grid
  - Not *3K*-random at all

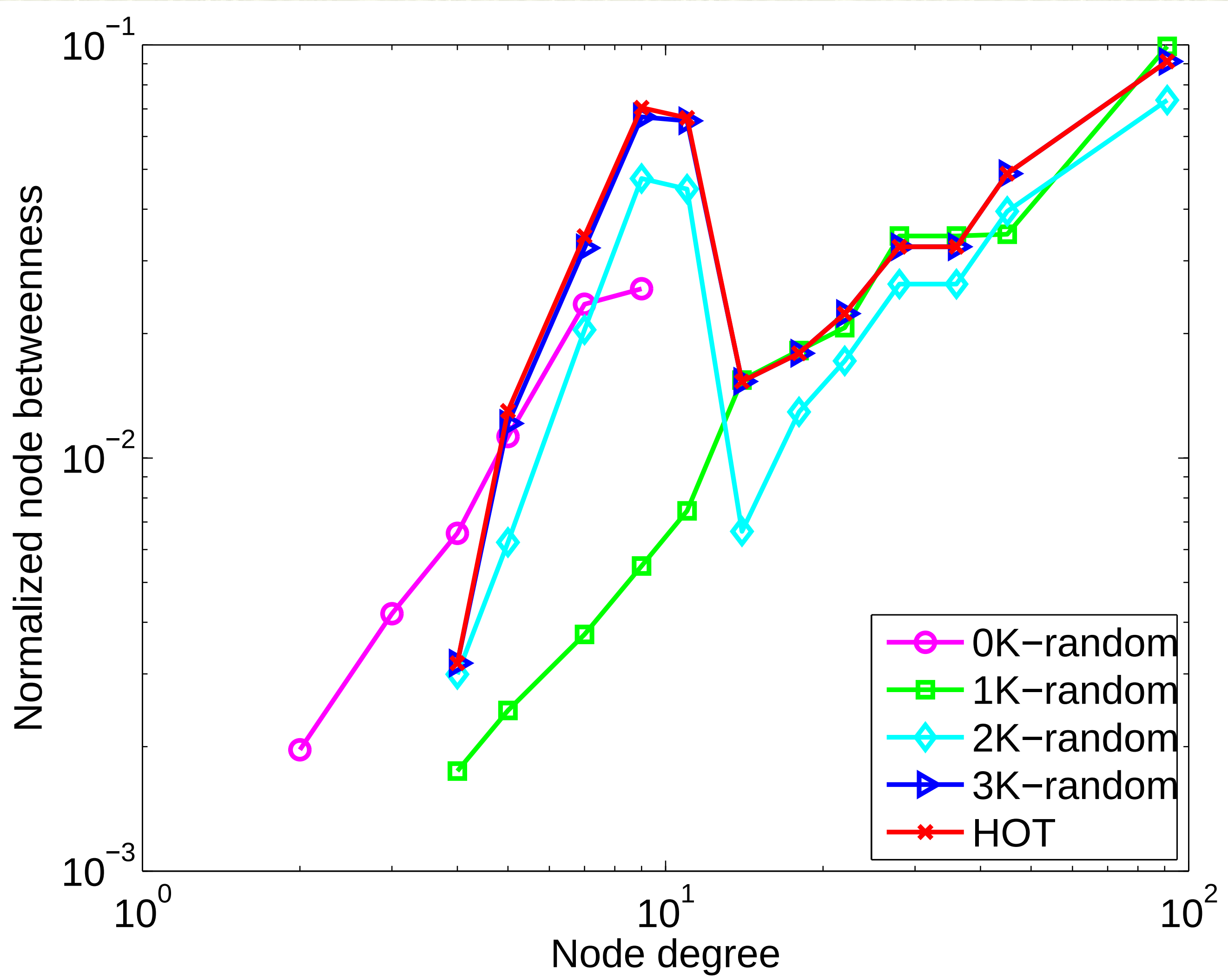
Methodology

- To show that a network is *dK*-random, it is sufficient to show that the difference between the *(d+1)K*-distribution in the network and in its *dK*-randomizations is statistically nonsignificant
  - We compute the statistical significance of motifs of size 4
- Just for fun, we also compute many other metrics and compare them between the network and its *dK*-randomizations
  - microscopic (degree distribution, correlations, clustering; motifs belong here, too)
  - mesoscopic (community structure)
  - macroscopic (distance and betweenness distributions)

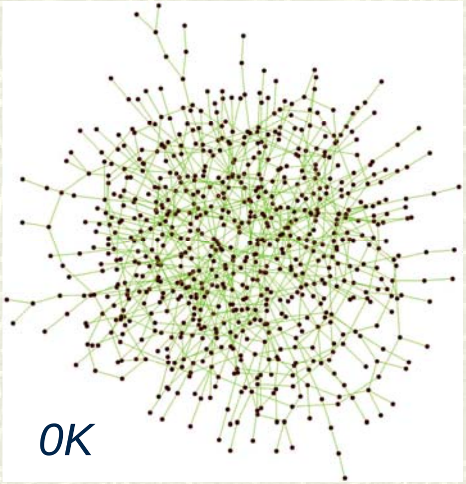
Internet AS-level (skitter): average neighbor degree

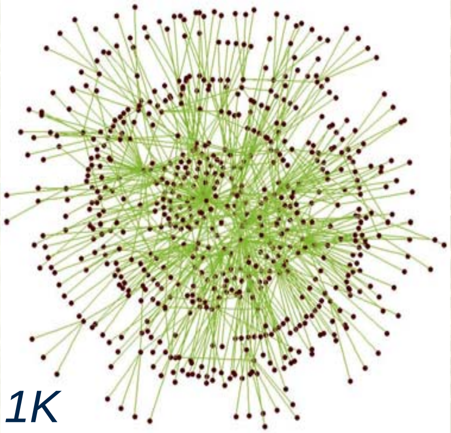


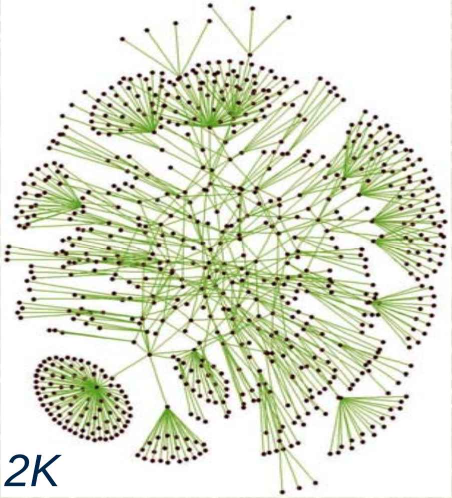
Internet "router"-level (HOT): degree-dependent betweenness

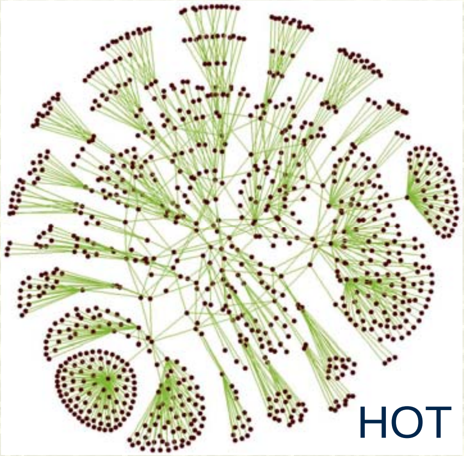


HOT *dK*-porn

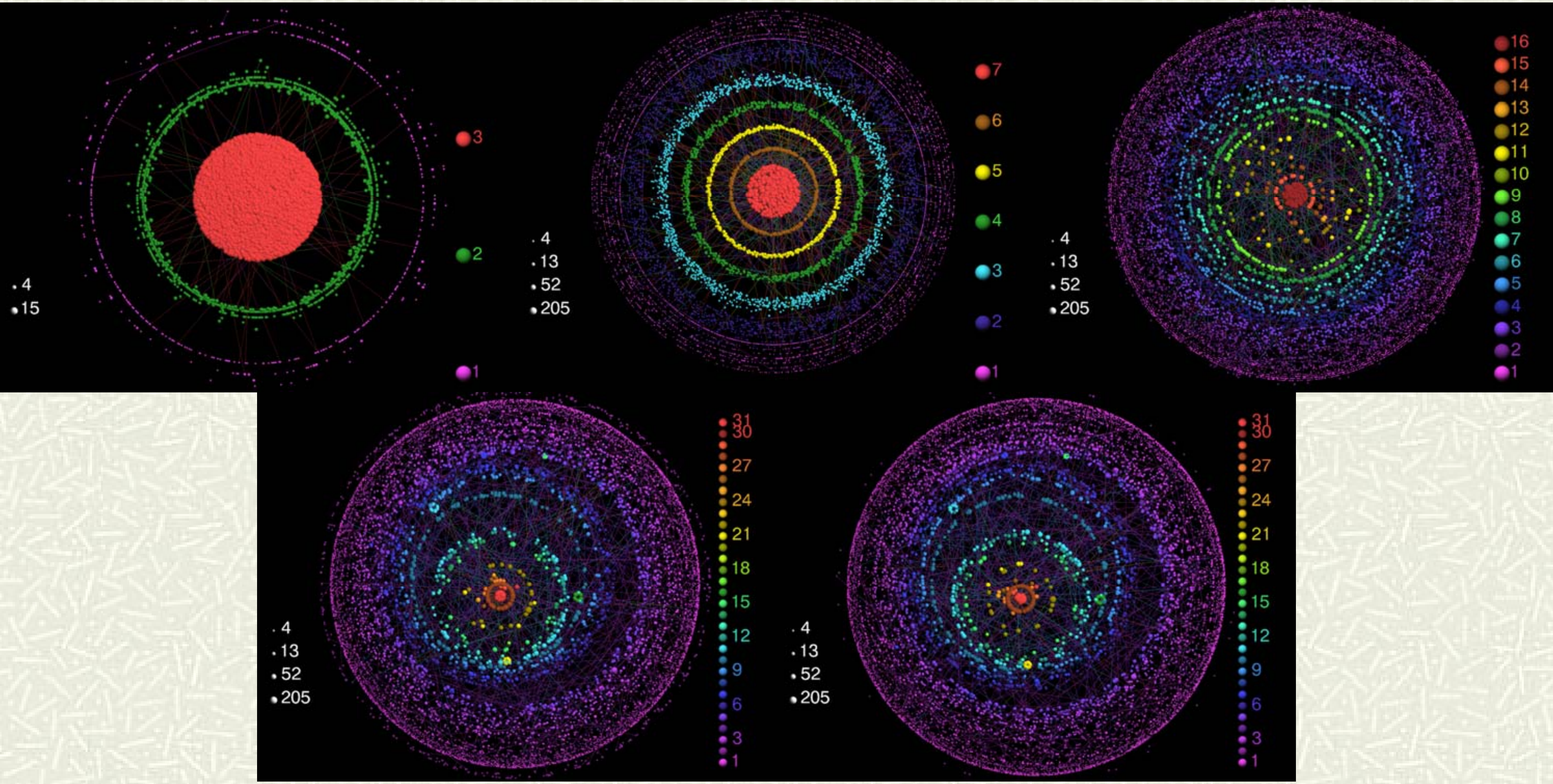






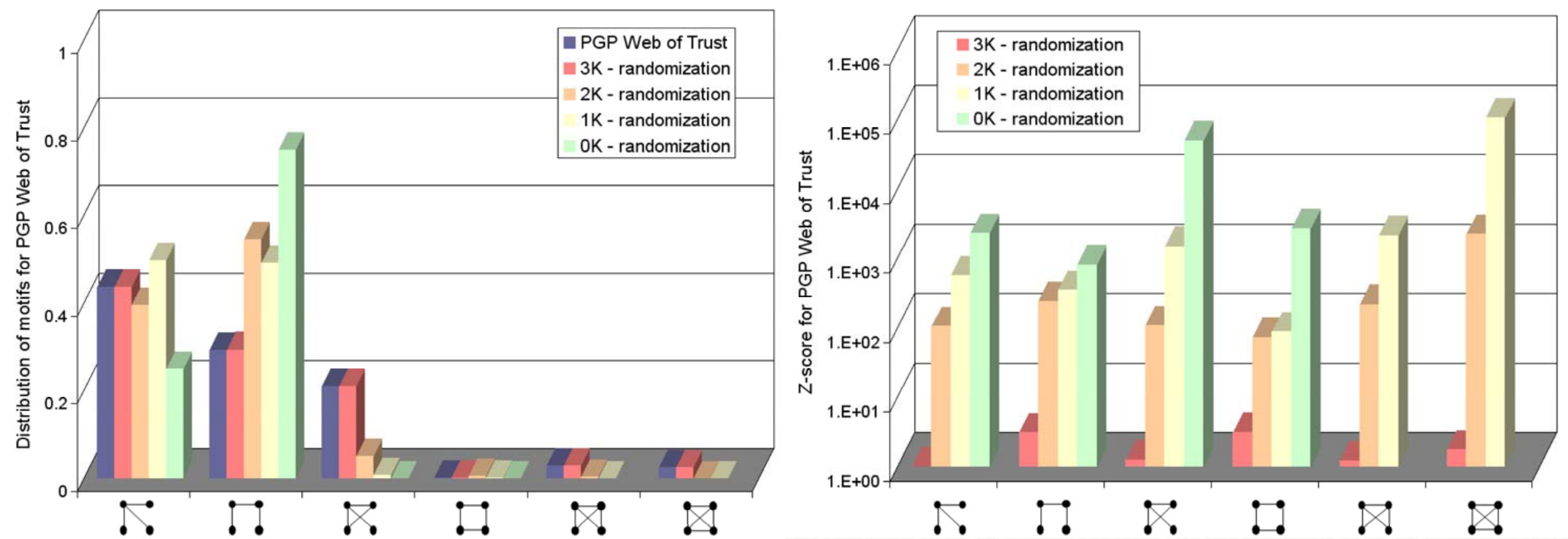


PGP *dK*-porn



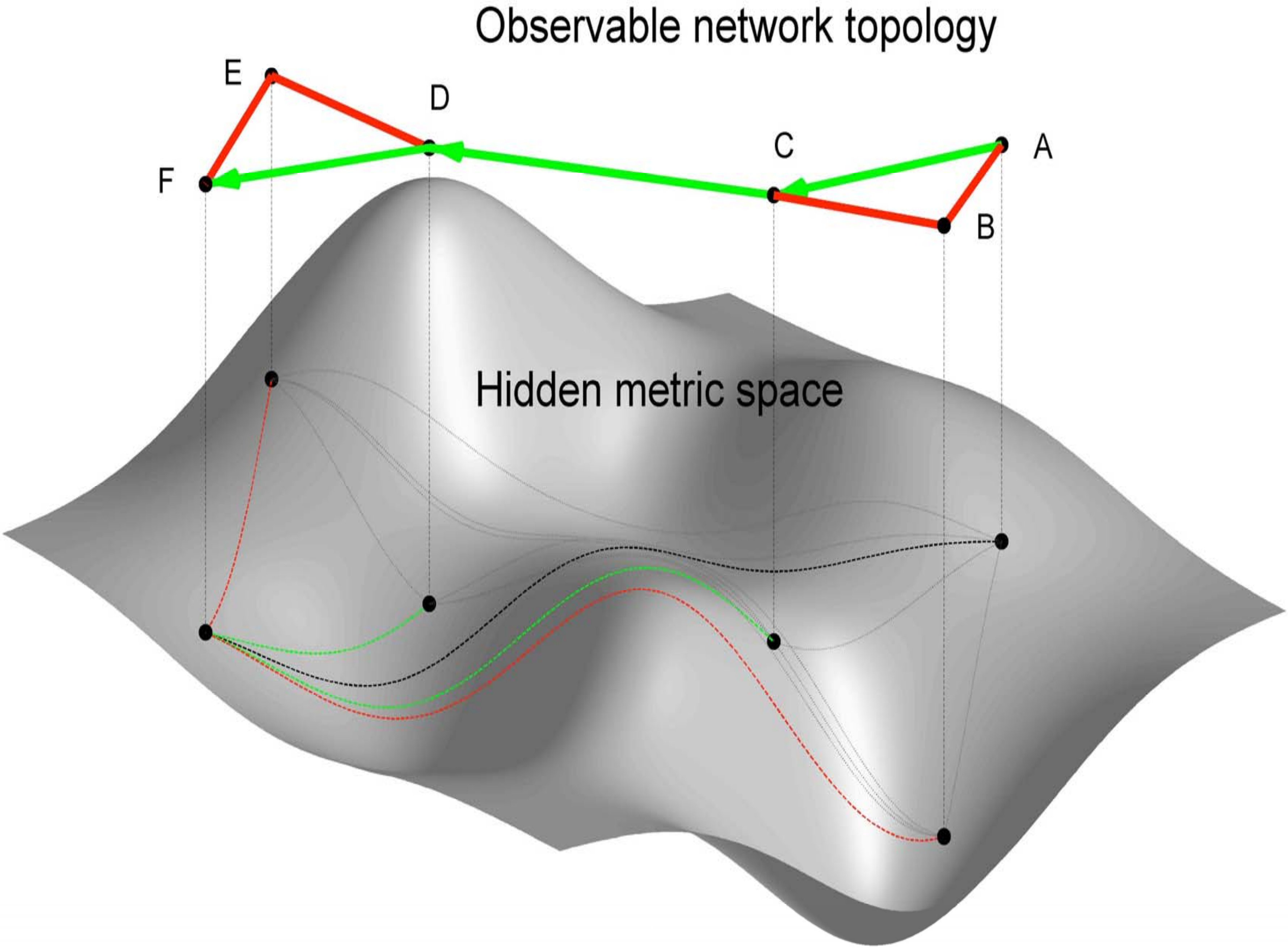
3
2
1
4
15
7
6
5
4
3
2
1
4
13
52
205
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
4
13
52
205
31
30
27
24
21
18
15
12
9
6
3
1
4
13
52
205
31
30
27
24
21
18
15
12
9
6
3
1
4
13
52
205

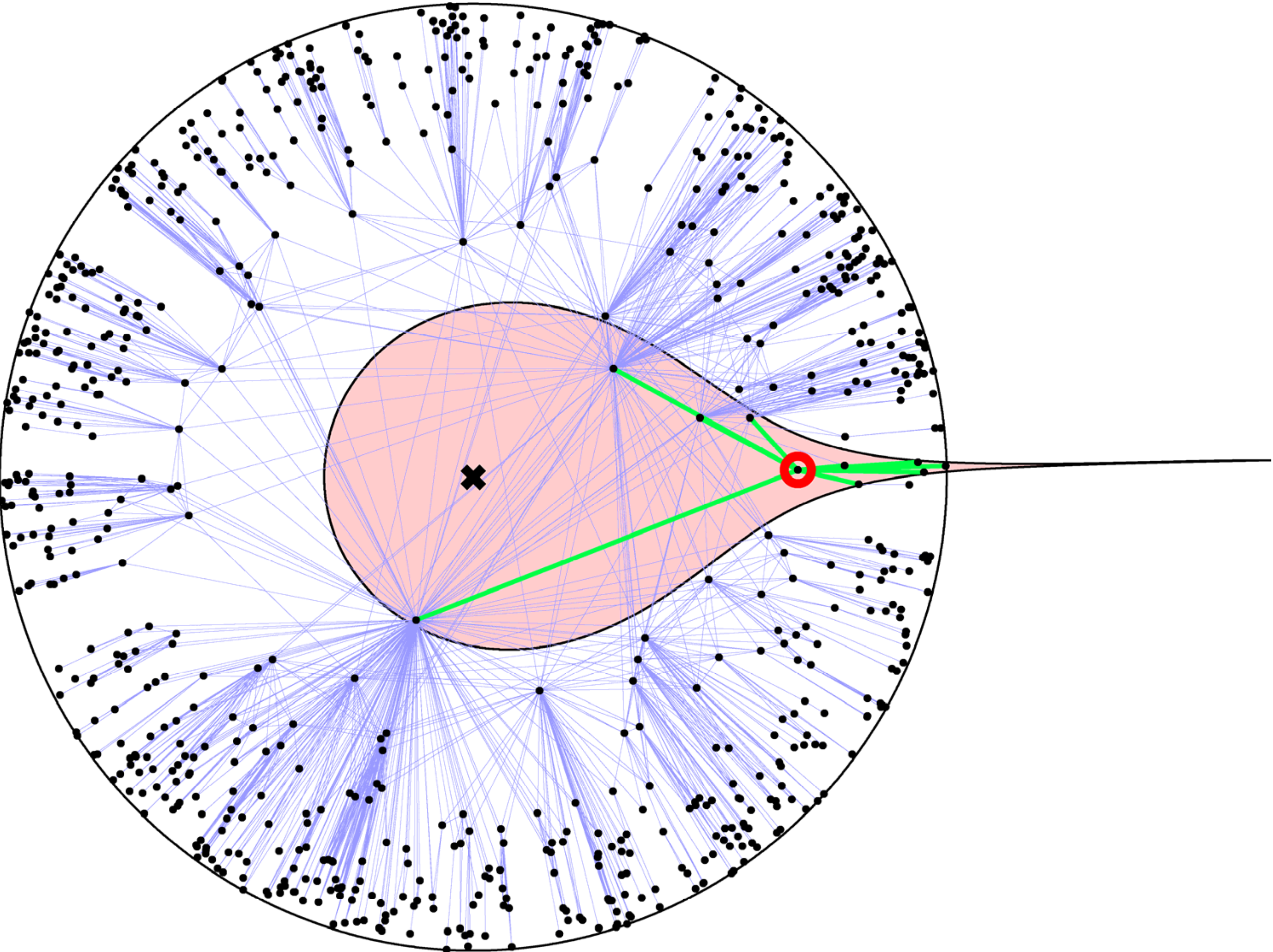
Web of trust (PGP): motifs of size 4



Outline

- Introduction
- *dK*-*
- *dK*-randomness of real networks
- Hidden hyperbolic metric spaces as an explanation
  - Hidden *metric* spaces and clustering
  - Hidden *hyperbolic* spaces and degree distribution
  - Degree distribution ∪ clustering ⊂ *3K*-distribution
- Conclusion


Observable network topology
E
D
C
A
F
B
Hidden metric space



Plausible explanation of ubiquitous *3K*-randomness

- The two main geometric properties of hidden spaces,
  - metric structure, and
  - negative curvature,

  explain the two main topological properties of complex networks,
  - strong clustering, and
  - power-law degree distributions
- Both are captured by the *3K*-distribution

Outline

- Introduction
- *dK*-*
- *dK*-randomness of real networks
- Hidden hyperbolic metric spaces as an explanation
- Conclusion
  - Take-home message
  - Implications
  - Speculations

Take-home message

- A majority of complex networks are *3K*-random at most

Implications

- Orbis is practically applicable not only to the Internet, but to many other networks as well
- Network evolution models and laws need not try to reproduce and explain the emergence of an endless list of metrics, but just the *3K*-distribution
  - Perhaps just the degree distribution and clustering
- Connection between network structure and function does not go via motifs
  - As soon as randomization basis is *3K*, all motifs are statistically non-significant

Speculations

- *Many* networks are *3K*-random, but not *all*, e.g., not the power grid. Why?
  - Unlikely because it is planar and spatially embedded
    - The airport network and the Internet are also spatially embedded, and the latter is even *1K*-random
  - More likely because it is a designed, engineered network, fully controlled by humans
  - As such it has lots of constraints, imposed by humans, that *dK*-series with low *d* cannot capture
  - It is good that we found a non-*3K*-random network, since it shows that "*d=3* is just too constraining" is not a satisfactory explanation of ubiquitous *3K*-randomness
- All self-evolving networks appear not to have any constraints other than hidden hyperbolic metric spaces