Popularity versus Similarity in Growing Networks

Fragiskos Papadopoulos
Cyprus University of Technology
M. Kitsak, M. Á. Serrano, M. Boguñá, and
Dmitri Krioukov
CAIDA/UCSD

SIAM CSE13, March 2013

Preferential Attachment (PA)

- *Popularity is attractive*
- If new connections in a growing network prefer popular (high-degree) nodes, then the network has a power-law distribution of node degrees
  - This result can be traced back to 1924 (Yule)

Issues with PA

- Zero clustering
- PA *per se* is ***impossible*** in real networks
  - It requires global knowledge of the network structure to be implemented
- The popularity preference should be exactly a linear function of the node degree
  - Otherwise, no power laws

One solution to these problems

- Mechanism:
  - New node selects an existing edge uniformly at random
  - And connects to its both ends
- Results:
  - No global intelligence
  - Effective linear preference
  - Power laws
  - Strong clustering
- Dorogovtsev *et al.*, PRE 63:062101, 2001

One problem with this solution

- It does not reflect reality
- It could not be validated against growth of real networks

No model that would:

- Be simple and universal (like PA)
  - Potentially describing (as a base line) evolution of many different networks
- Yield graphs with observable properties
  - Power laws, strong clustering, to start with
  - But many other properties as well
- Not require any global intelligence
- Be ***validated***

Validation of growth mechanism

- State of the art
  - Here is my new model
  - The graphs that it produces have power laws!
  - ~~And strong clustering!!~~
  - And even X!!!
- Almost never the growth mechanism is validated ***directly***
- PA *was* validated directly for many networks, because it is so simple

Paradox with PA validation

- Dilemma
  - PA was validated
  - But PA is impossible
- Possible resolution
  - PA is an emergent phenomenon
  - A consequence of some other underlying processes

Popularity versus Similarity

- Intuition
  - I (new node) connect to you (existing node) not only if you are popular (like Google or Facebook), but also if you are similar to me (like Tartini or free soloing) — homophily
- Mechanism
  - New connections are formed by trade-off optimization between popularity and similarity

Mechanism (growth algorithm)

- Nodes t are introduced one by one
  - $t = 1, 2, 3, \ldots$
- Measure of popularity
  - Node's birth time t
- Measure of similarity
  - Upon its birth, node t gets positioned at a random coordinate θ_t in a "similarity" space
  - The similarity space is a circle
  - θ is random variable uniformly distributed on $[0,2\pi]$
  - Measure of similarity between t and s is $\theta_{st} = |\theta_s - \theta_t|$

Mechanism (contd.)

- New connections
  - New node t connects to m existing nodes s, $s < t$, minimizing $s\theta_{st}$
  - That is, maximizing the product between popularity and similarity

New node *t* connects to *m* existing nodes *s* that minimize

$$s\theta_{st}$$

$$st\frac{\theta_{st}}{2}$$

$$\ln\left(st\frac{\theta_{st}}{2}\right)$$

$$= r_s + r_t + \ln\frac{\theta_{st}}{2}$$

$$\approx x_{st}$$ — the *hyperbolic* distance between *s* and *t*

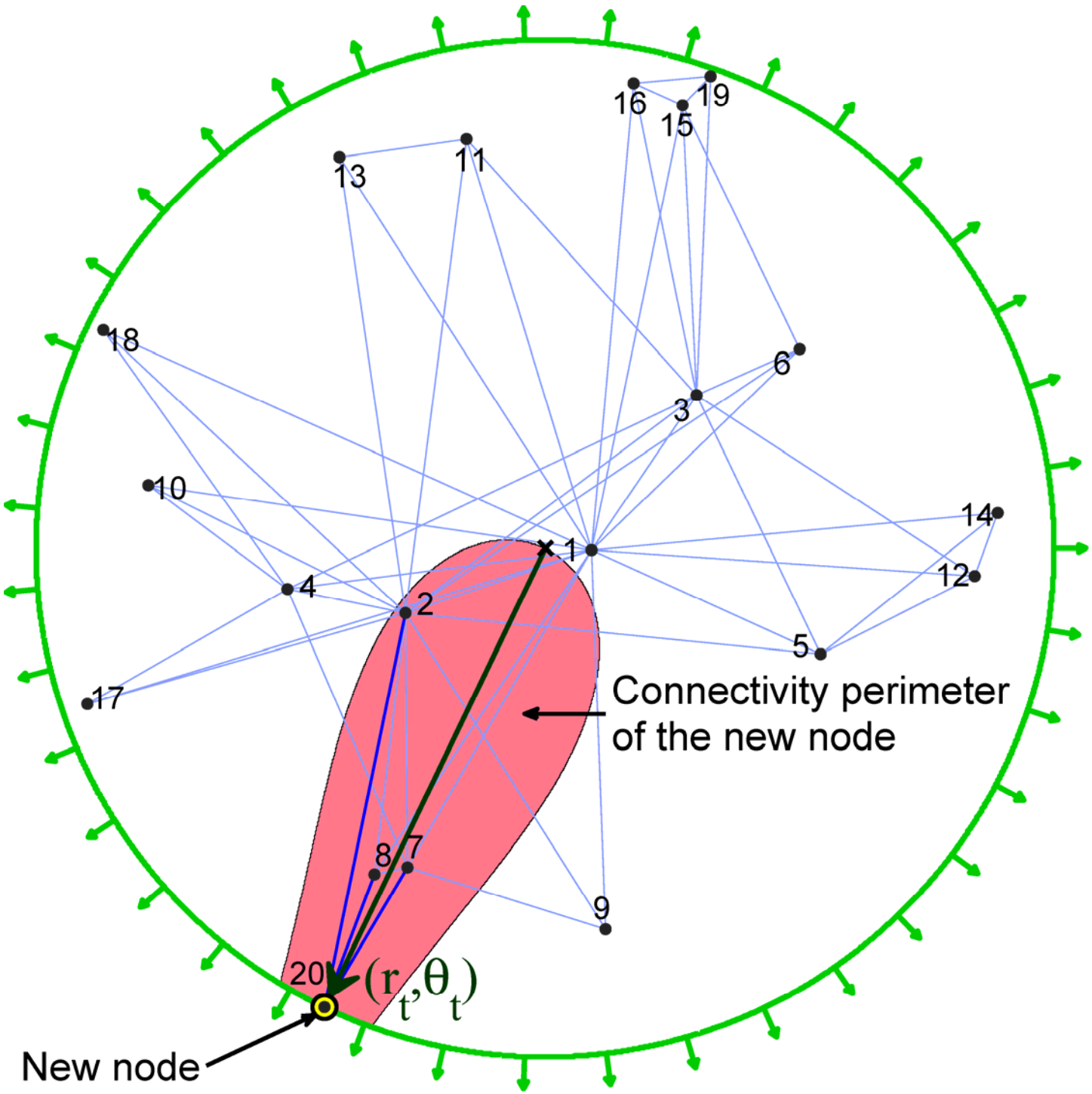
New nodes connects to *m* hyperbolically closest nodes

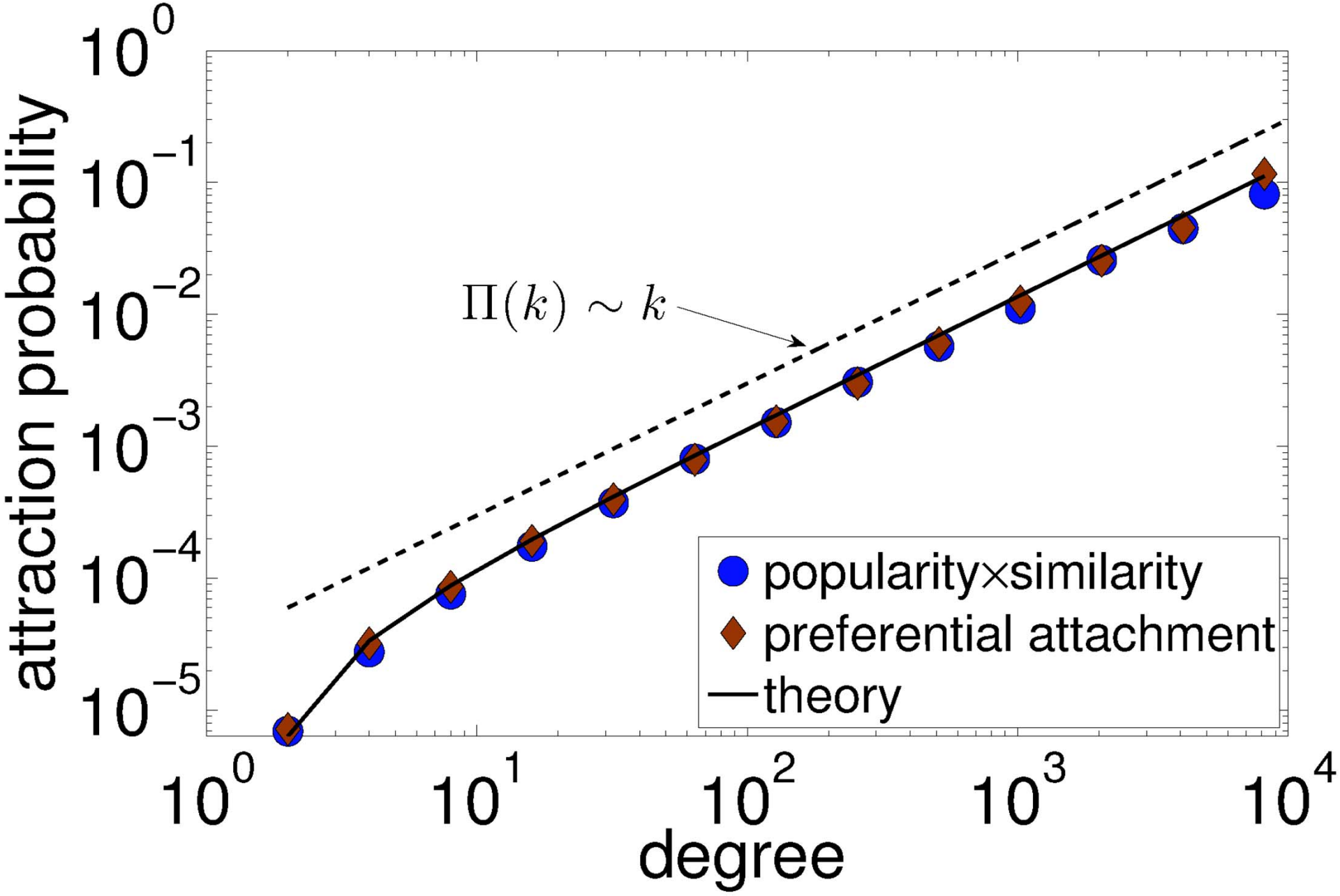
New nodes connects to m hyperbolically closest nodes

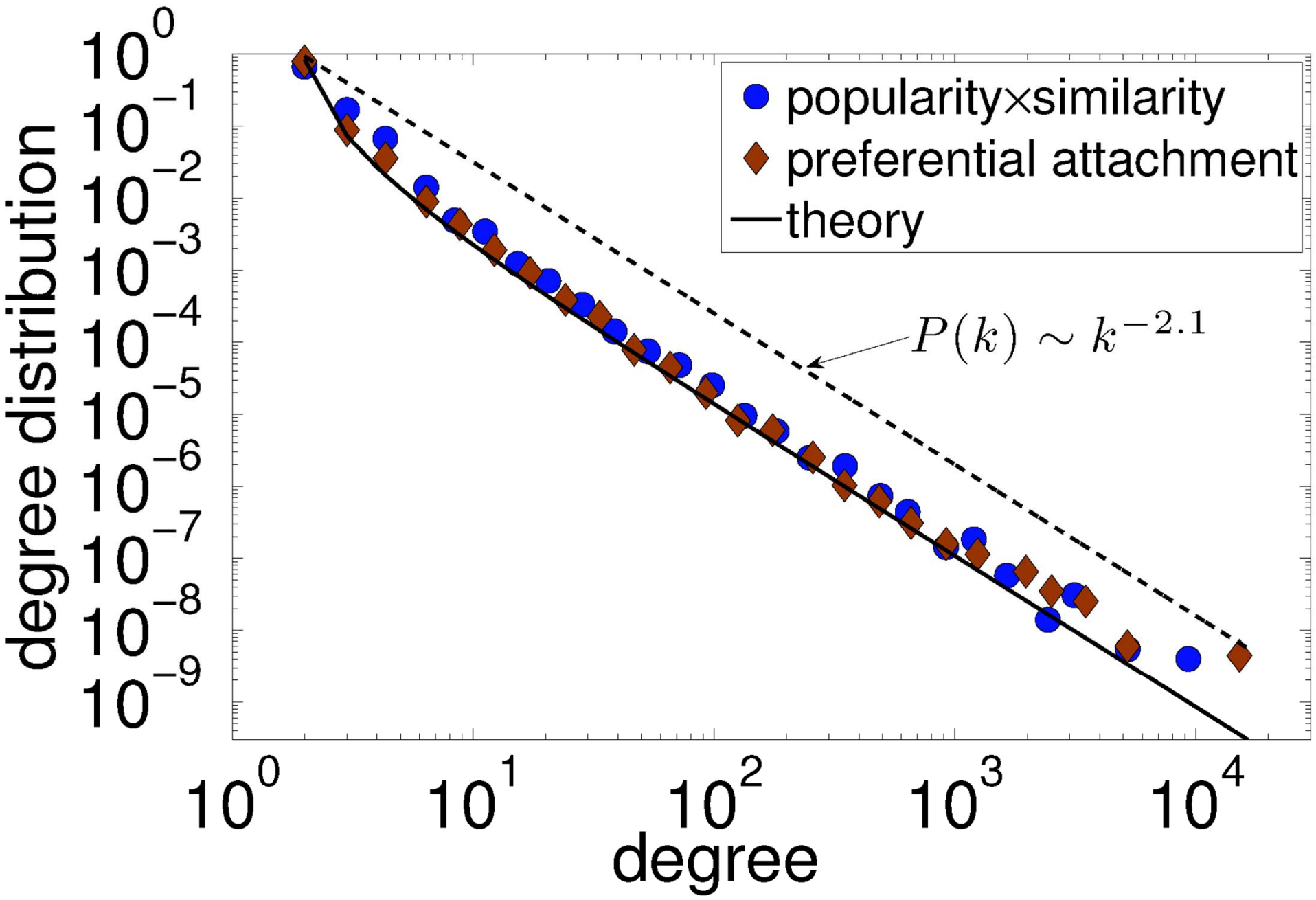
The expected distance to the m'th closest node from t is

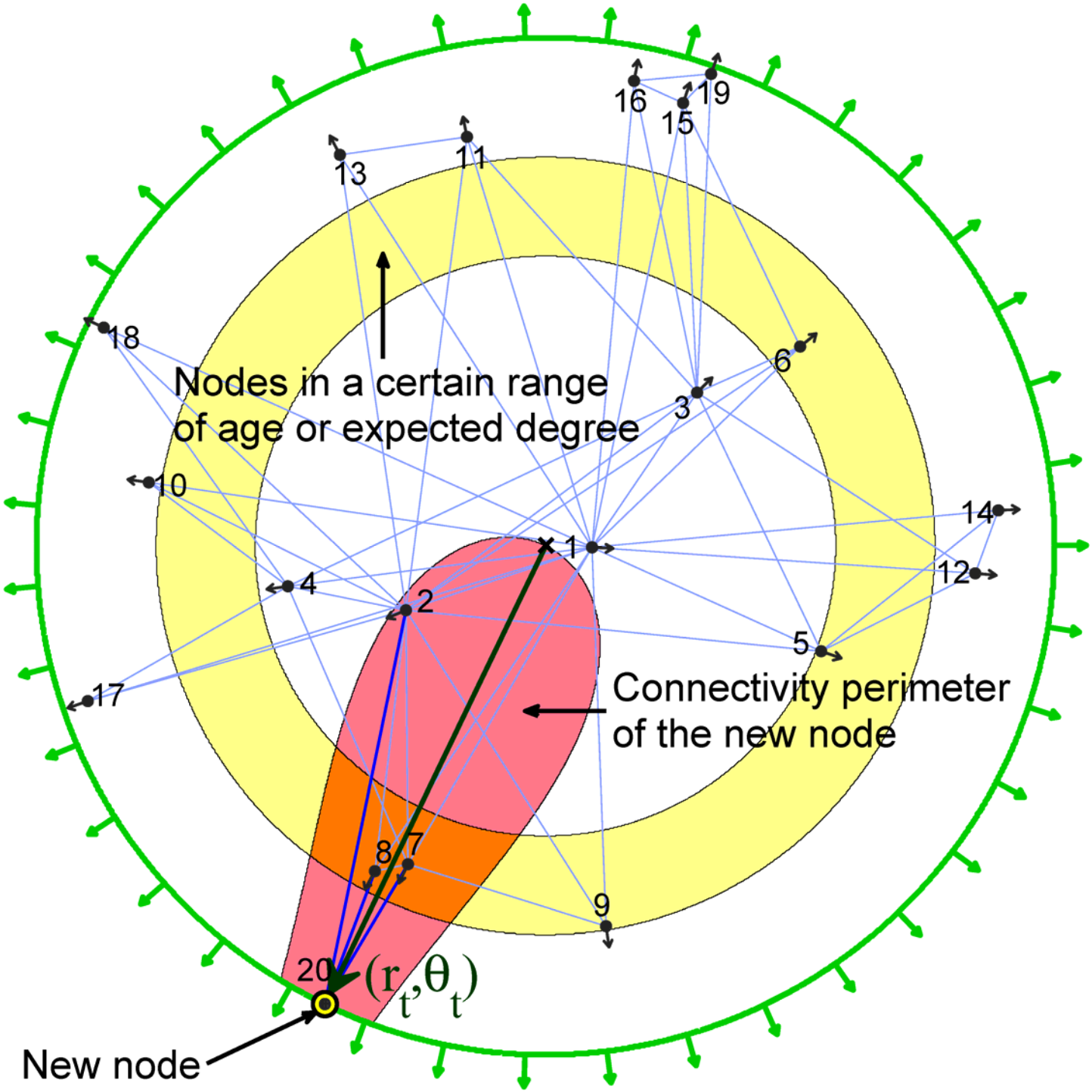
$$R_t = \ln \frac{\pi m t}{2\left(1 - \frac{1}{t}\right)} \approx r_t + \ln \frac{\pi m}{2} \approx r_t$$

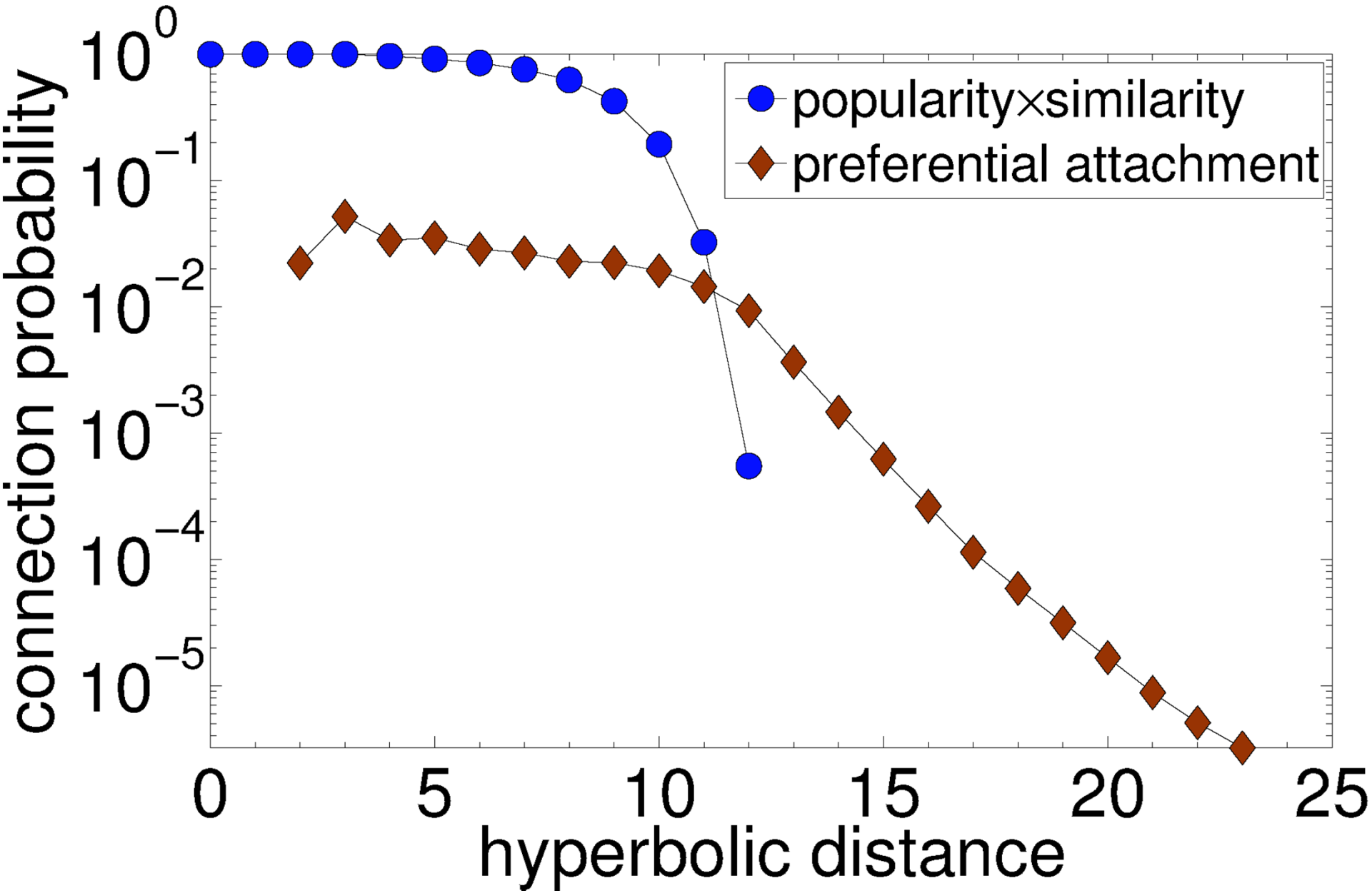
New node t is located at radial coordinate $r_t \sim \ln t$,
and connects to all nodes within distance $R_t \sim r_t$

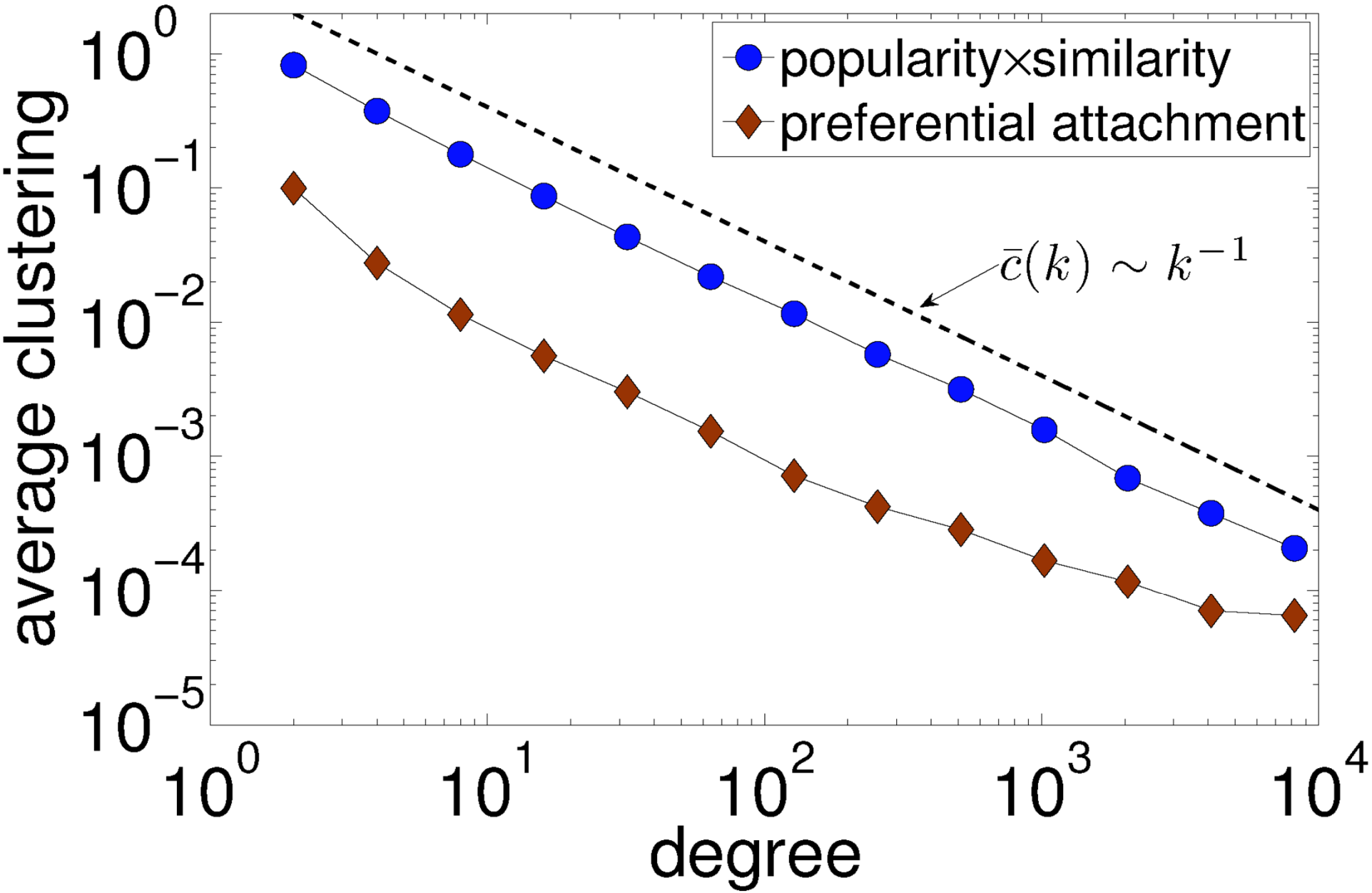

16
19
15
13
11
18
6
3
10
14
1
12
4
2
5
17
Connectivity perimeter
of the new node
8
7
9
20
(r_t, θ_t)
New node


attraction probability
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
$\Pi(k) \sim k$
popularity×similarity
preferential attachment
theory
10^0
10^1
10^2
10^3
10^4
degree


popularity×similarity
preferential attachment
theory
$P(k) \sim k^{-2.1}$
degree distribution
degree
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^0
10^1
10^2
10^3
10^4

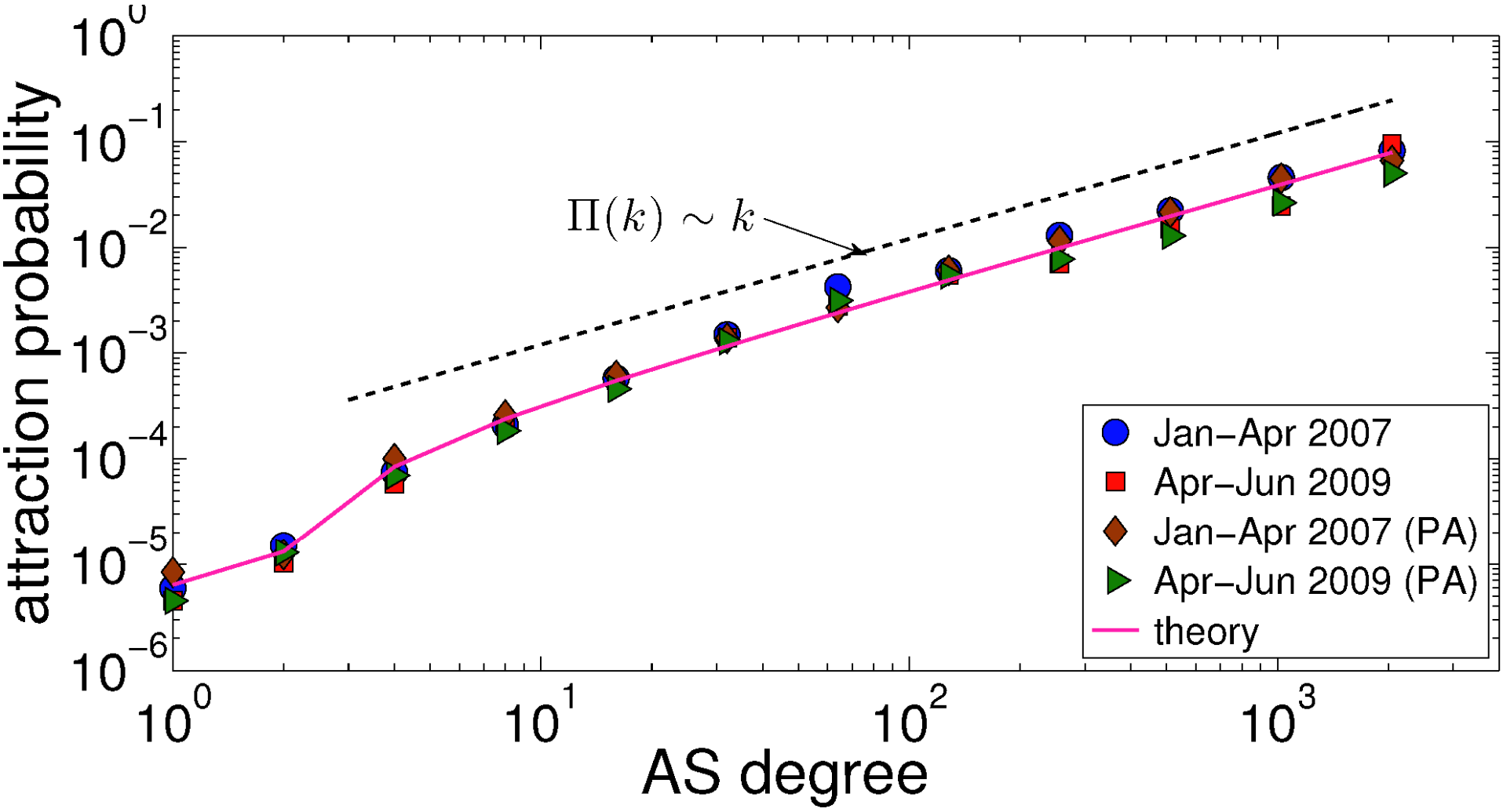

16
19
15
13
11
18
Nodes in a certain range
of age or expected degree
6
3
10
14
1
12
4
2
5
17
Connectivity perimeter
of the new node
8
7
9
20
(r_t, θ_t)
New node

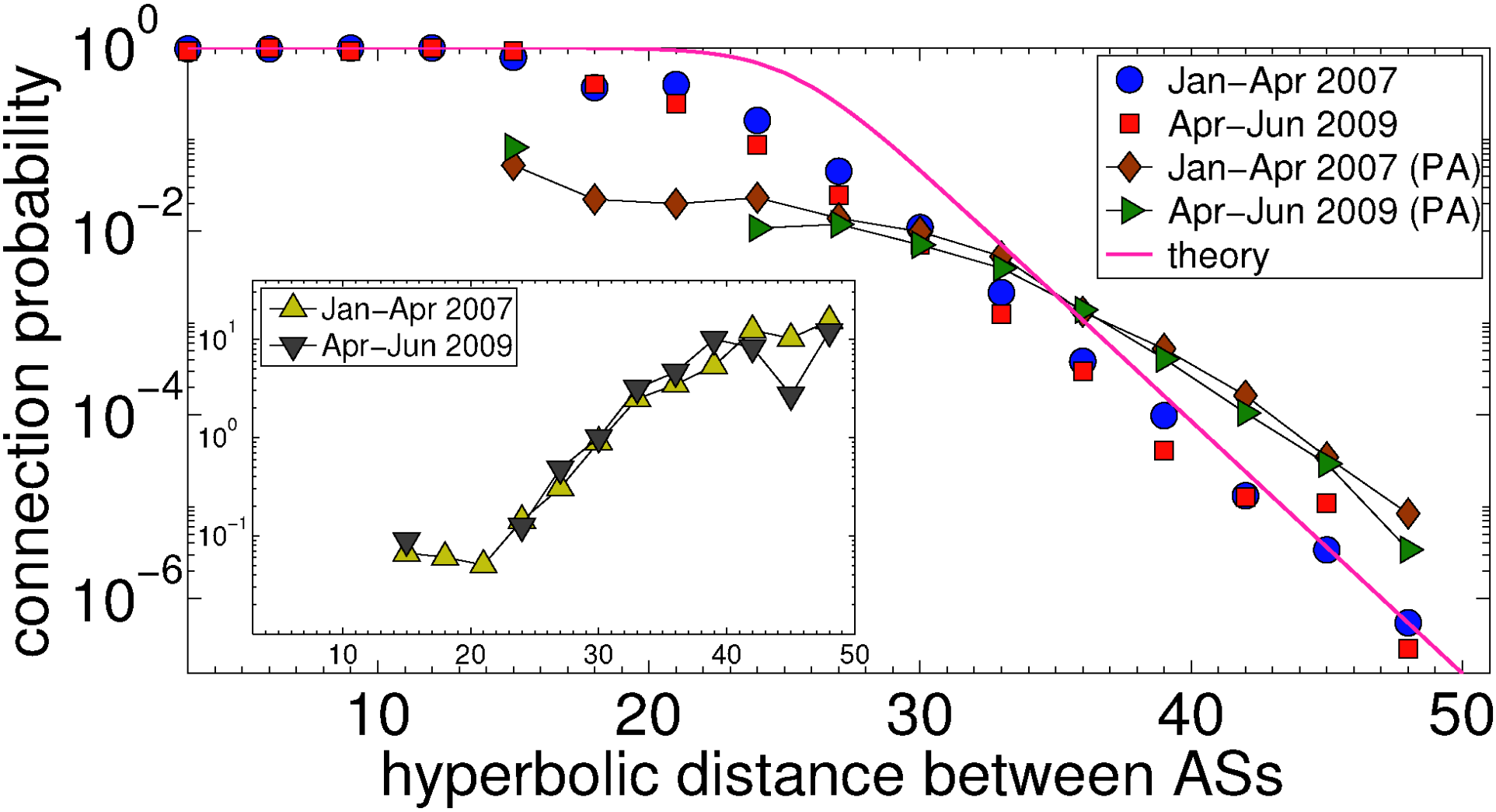

connection probability
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
popularity×similarity
preferential attachment
0
5
10
15
20
25
hyperbolic distance

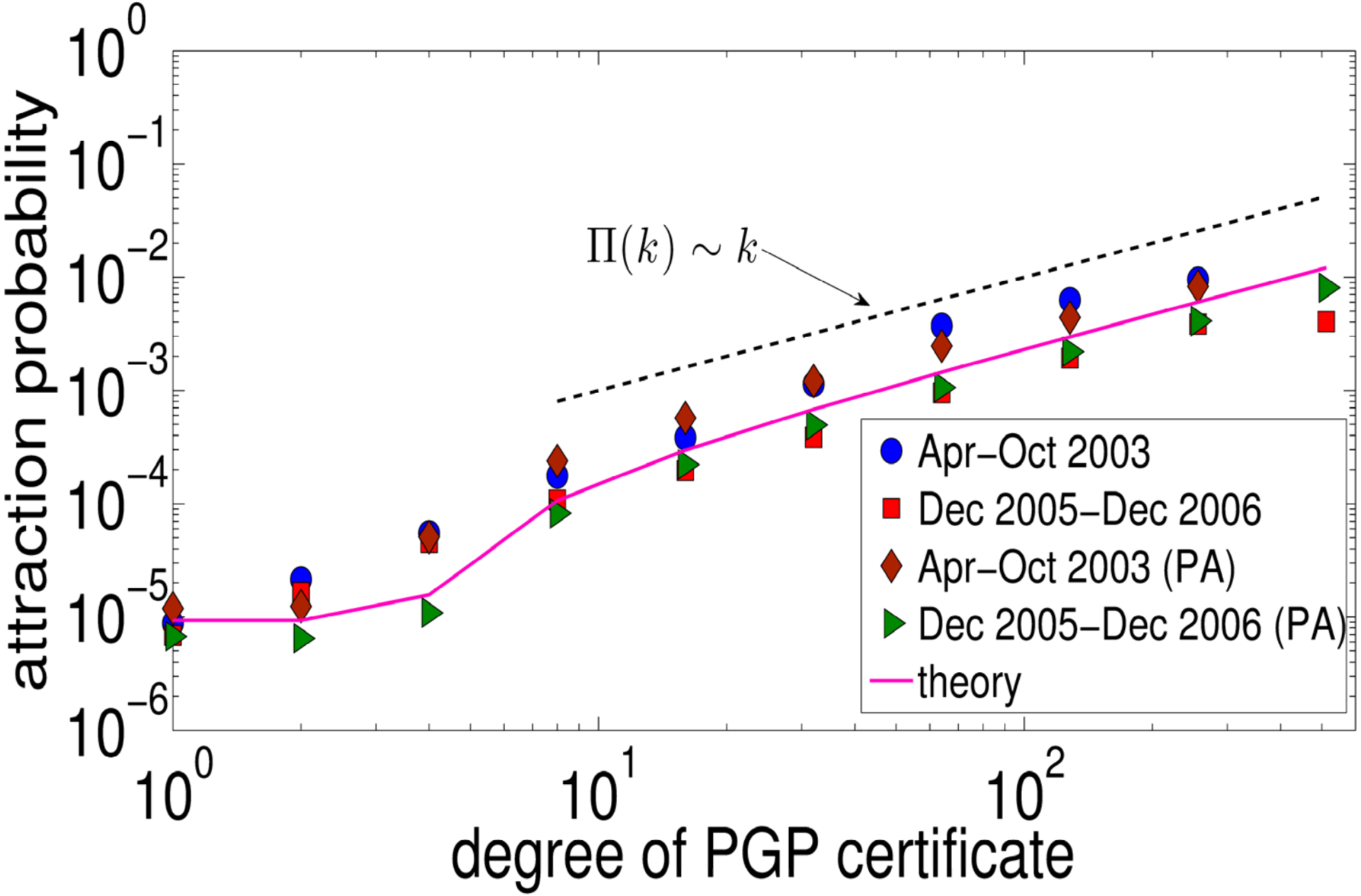

popularity×similarity
preferential attachment
$\bar{c}(k) \sim k^{-1}$
average clustering
degree
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^0
10^1
10^2
10^3
10^4

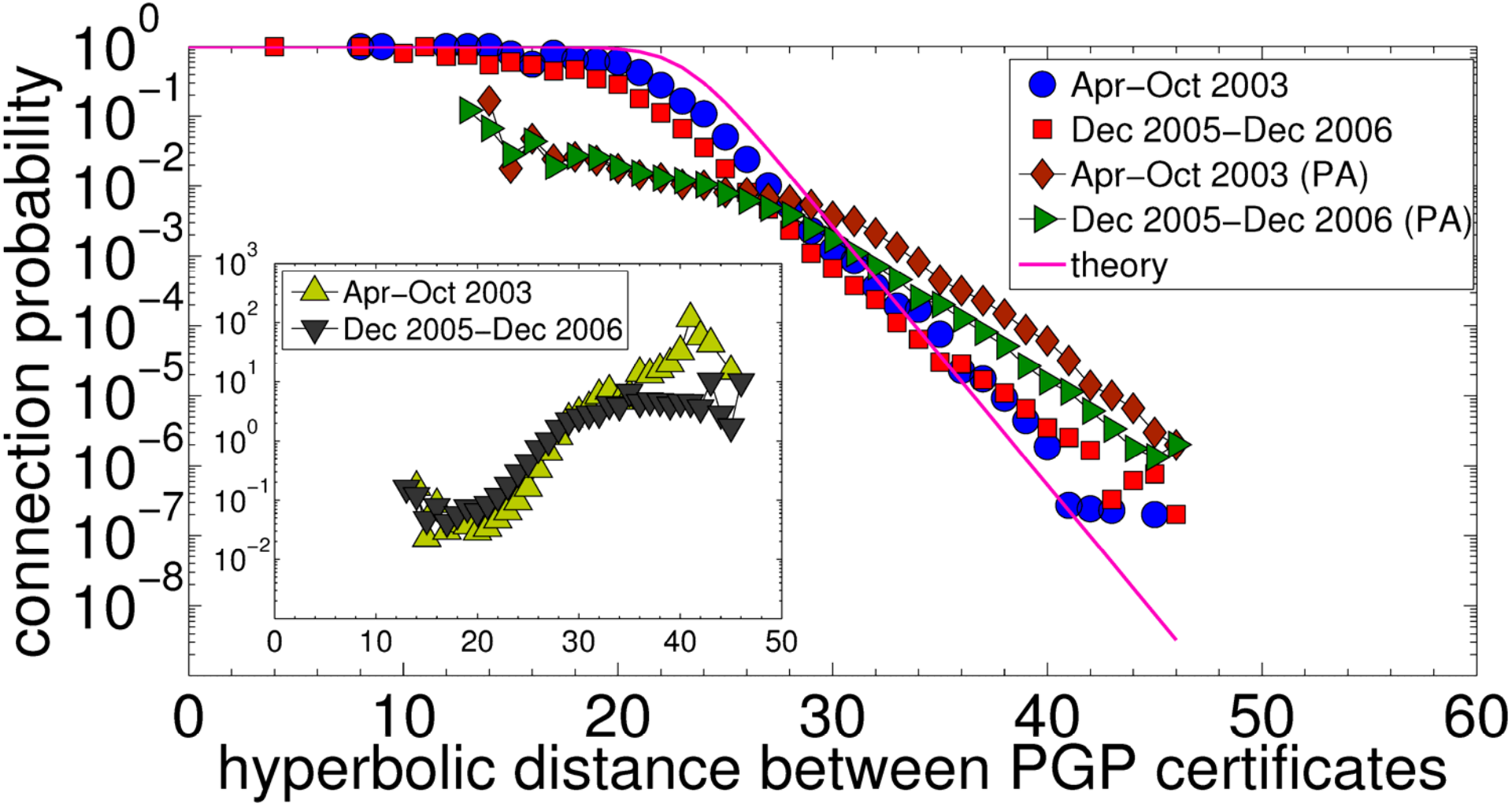
Validation

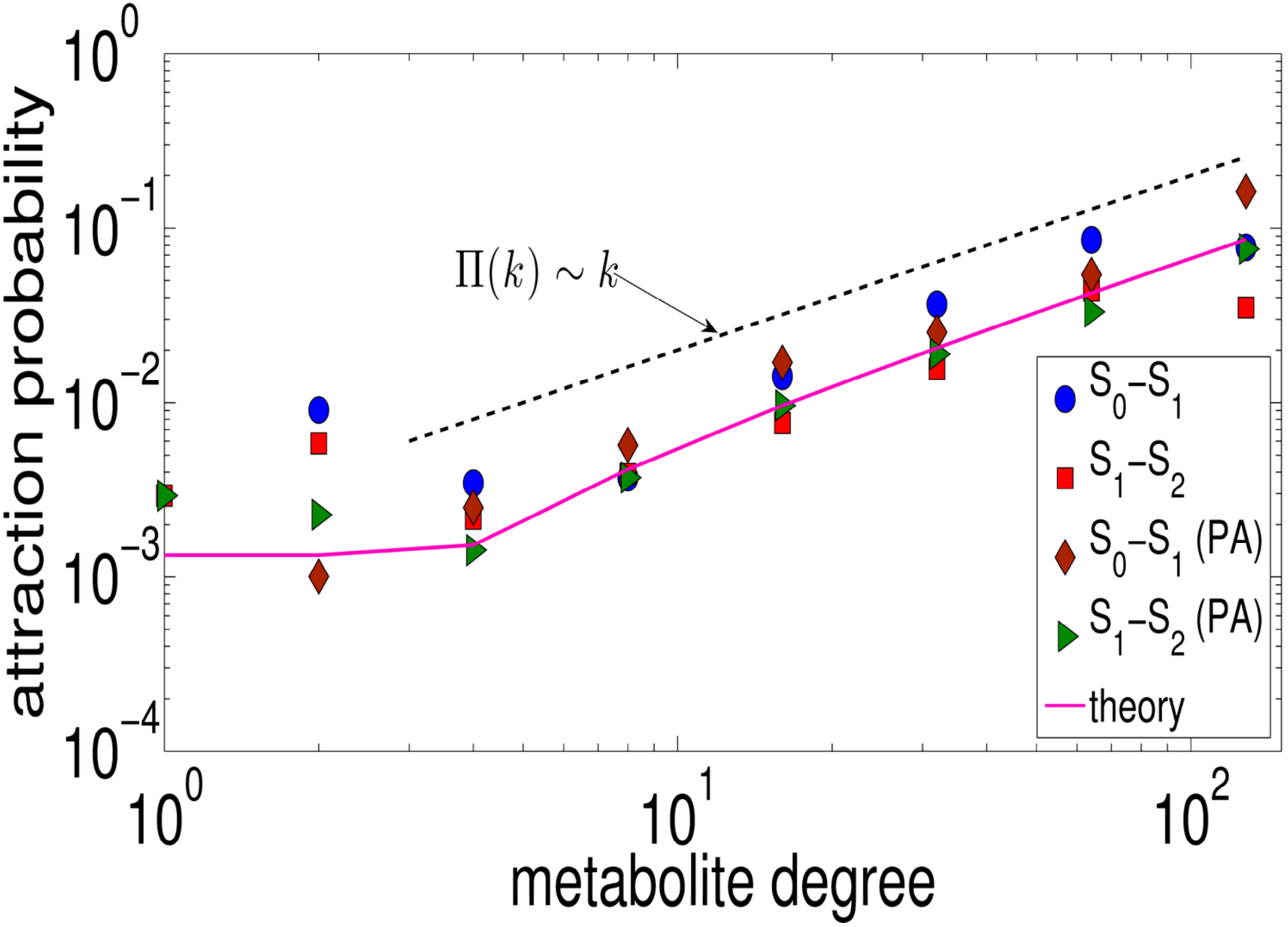
- Take a series of historical snapshots of a real network
- Infer angular/similarity coordinates for each node
- Test if the probability of new connections follows the model theoretical prediction

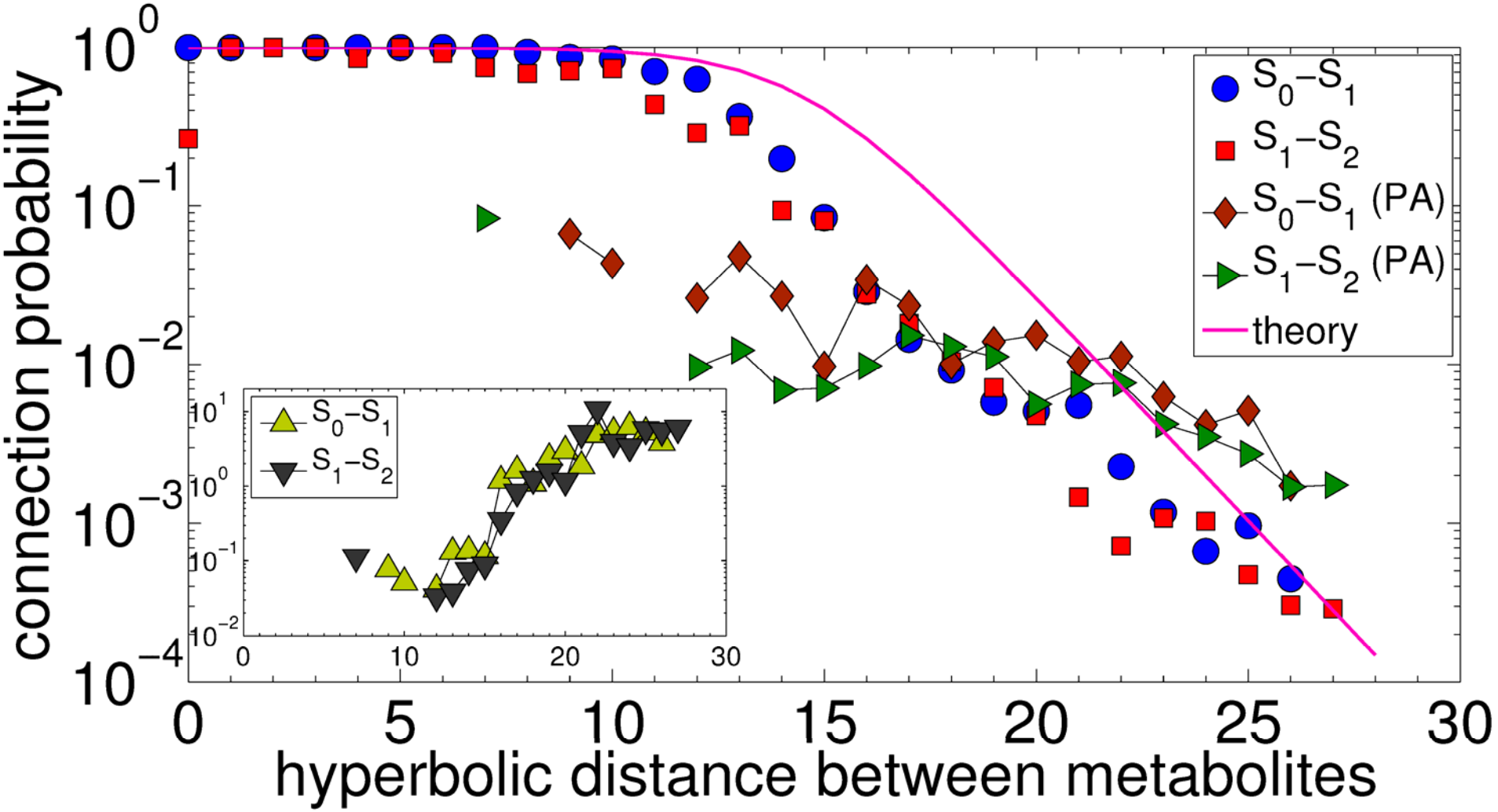

attraction probability
AS degree
$\Pi(k) \sim k$
Jan–Apr 2007
Apr–Jun 2009
Jan–Apr 2007 (PA)
Apr–Jun 2009 (PA)
theory


connection probability
hyperbolic distance between ASs
Jan–Apr 2007
Apr–Jun 2009
Jan–Apr 2007 (PA)
Apr–Jun 2009 (PA)
theory
Jan–Apr 2007
Apr–Jun 2009


attraction probability
degree of PGP certificate
$\Pi(k) \sim k$
Apr–Oct 2003
Dec 2005–Dec 2006
Apr–Oct 2003 (PA)
Dec 2005–Dec 2006 (PA)
theory


connection probability
hyperbolic distance between PGP certificates
Apr–Oct 2003
Dec 2005–Dec 2006
Apr–Oct 2003 (PA)
Dec 2005–Dec 2006 (PA)
theory
Apr–Oct 2003
Dec 2005–Dec 2006


attraction probability
metabolite degree
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^0
10^1
10^2
$\Pi(k) \sim k$
S_0–S_1
S_1–S_2
S_0–S_1 (PA)
S_1–S_2 (PA)
theory


connection probability
hyperbolic distance between metabolites
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
5
10
15
20
25
30
S_0–S_1
S_1–S_2
S_0–S_1 (PA)
S_1–S_2 (PA)
theory
10^1
10^0
10^{-1}
10^{-2}
0
10
20
30
S_0–S_1
S_1–S_2

Popularity×similarity optimization

- Explains PA as an emergent phenomenon
- Resolves all major issues with PA
- Generates graphs similar to real networks across many vital metrics
- *Directly* validates against some real networks
  - Technological (Internet)
  - Social (web of trust)
  - Biological (metabolic)
  - **Universe**

PSO compared to PA

- PA just ignores similarity (or hidden space), which leads to severe aberrations
  - Probability of similar (spatially close) connections is badly underestimated
  - Probability of dissimilar (spatially distant) connections is badly overestimated
- If the connection probability is correctly estimated, then one immediate application is link prediction
- PSO-based missing link prediction in the Internet outperforms all popular methods

Bottom line

- PA is a degenerate (infinite-temperature) regime with similarity/homophily factors reduced to nothing but noise
- If we take these factors into account, then
  - We can predict large-scale growth dynamics of real networks with a remarkable accuracy
  - This growth dynamics has seemingly nothing to do with PA (optimization vs. randomness)
  - Yet if one looks only at degree-based statistics, there is no difference

- F. Papadopoulos, M. Kitsak, M. Á. Serrano, M. Boguñá, and D. Krioukov,
  Popularity versus Similarity in Growing Networks,
  Nature, v.489, p.537, 2012