Popularity versus Similarity in Growing Networks

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

Caltech, February 2012

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

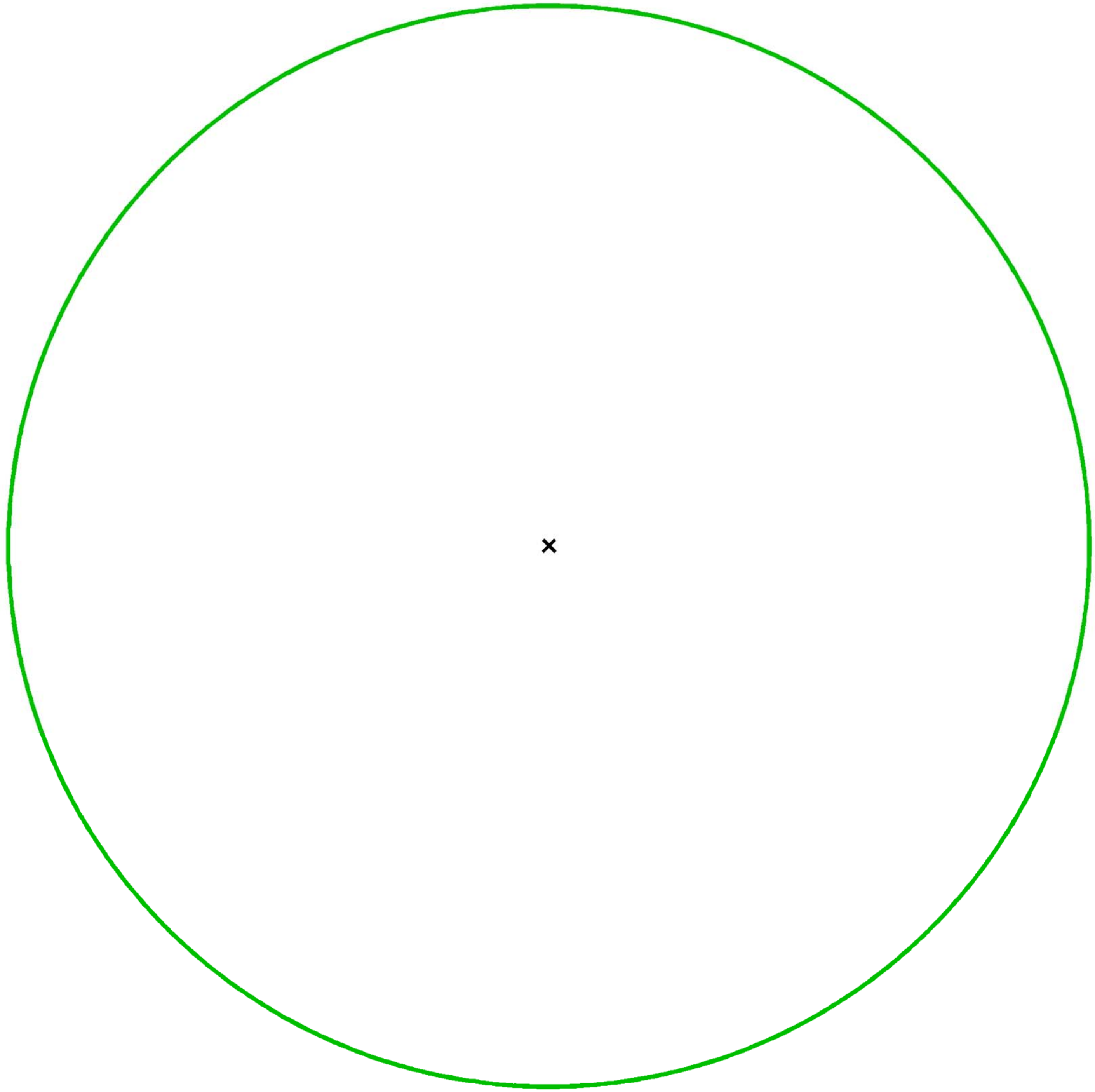
- Intuition
  - I (new node) connect to you (existing node) not only if you are popular (like Google or Facebook), but also if you are similar to me (like Tartini or free soloing) — homophily
- Mechanism
  - New connections are formed by trade-off optimization between popularity and similarity

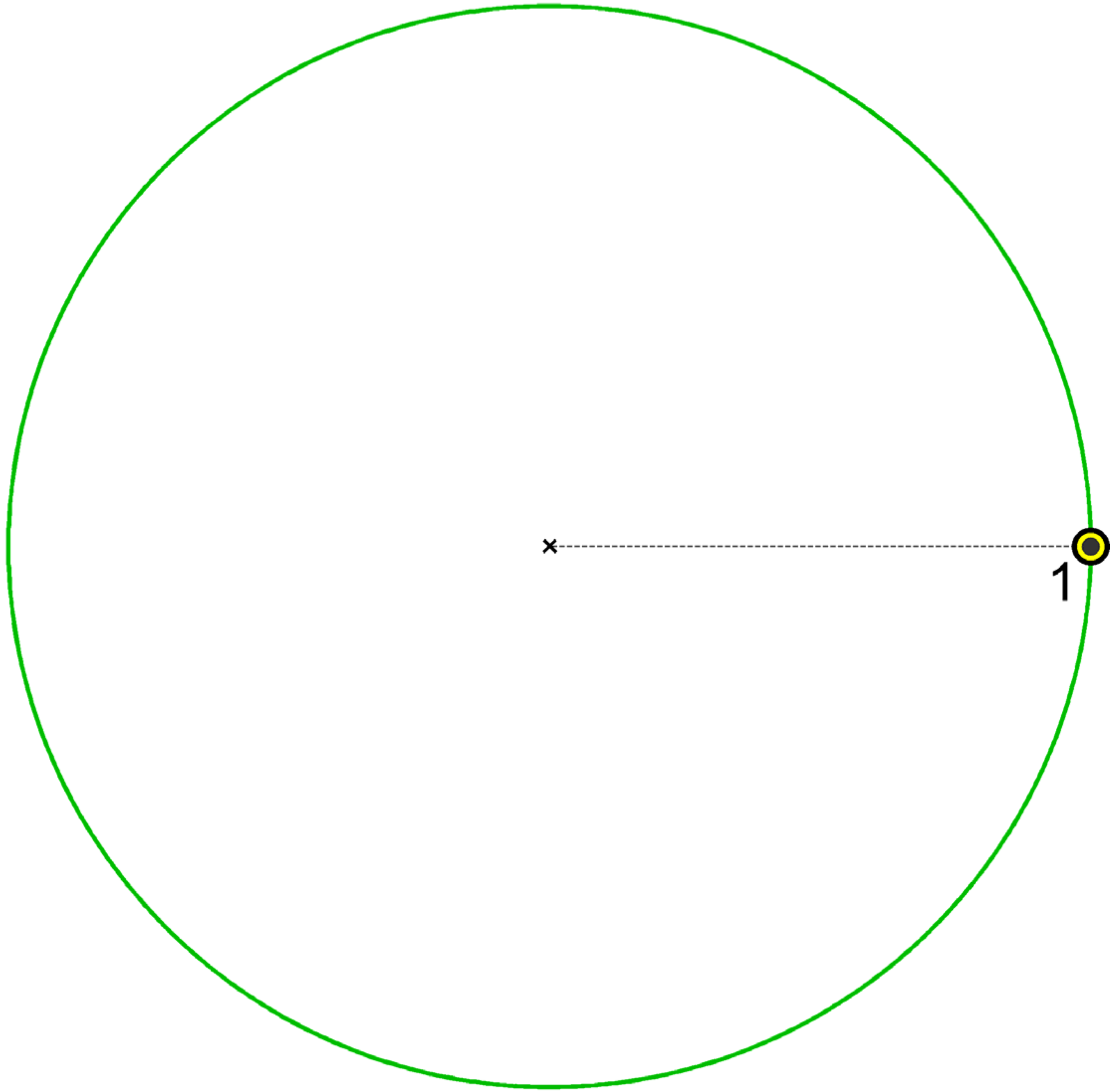
Mechanism (growth algorithm)

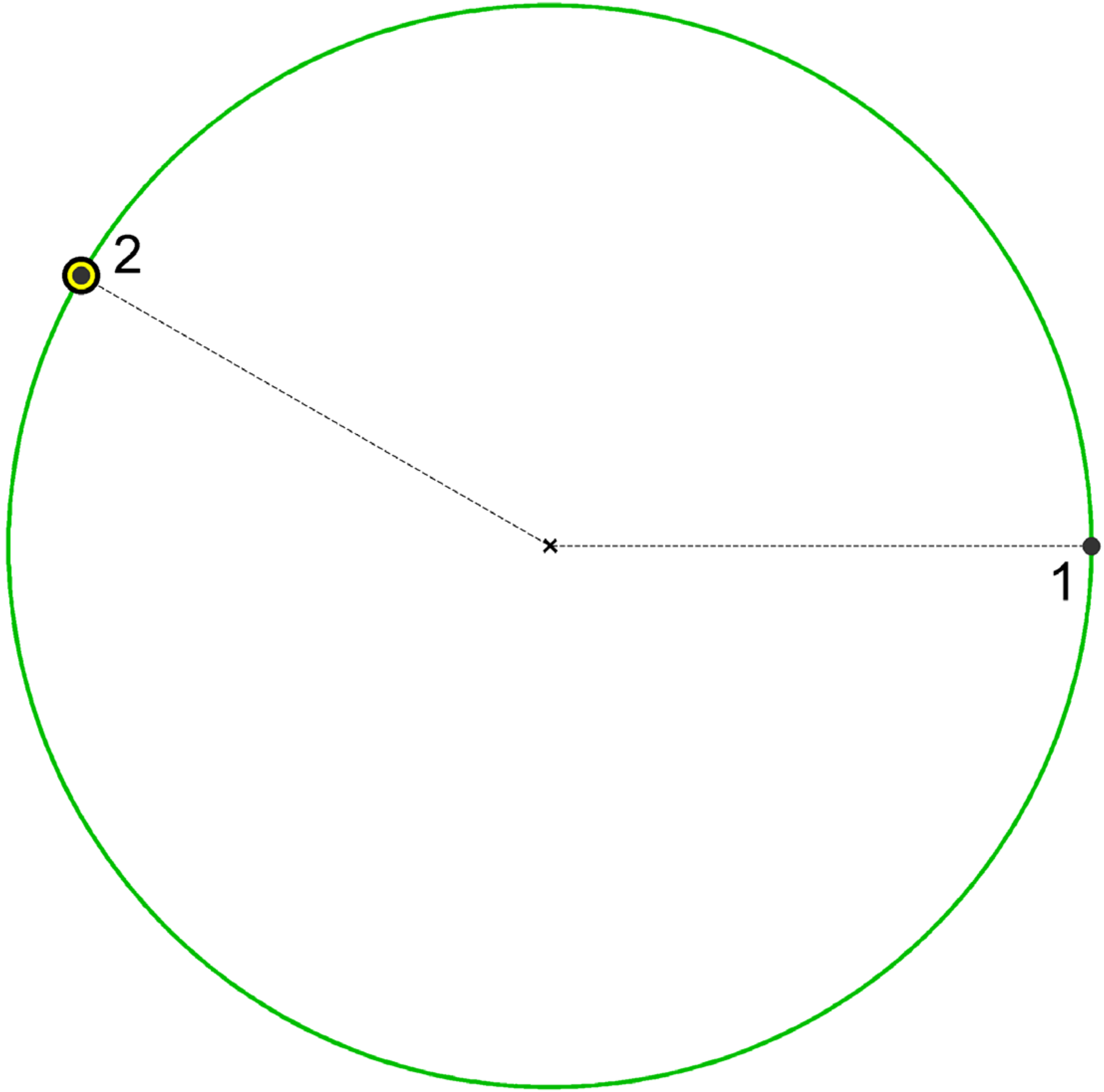
- Nodes *t* are introduced one by one
  - t @ 1, 2, 3, …
- Measure of popularity
  - Node's birth time *t*
- Measure of similarity
  - Upon its birth, node *t* gets positioned at a random coordinate θ_t in a "similarity" space
  - The similarity space is a circle
  - θ is random variable uniformly distributed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

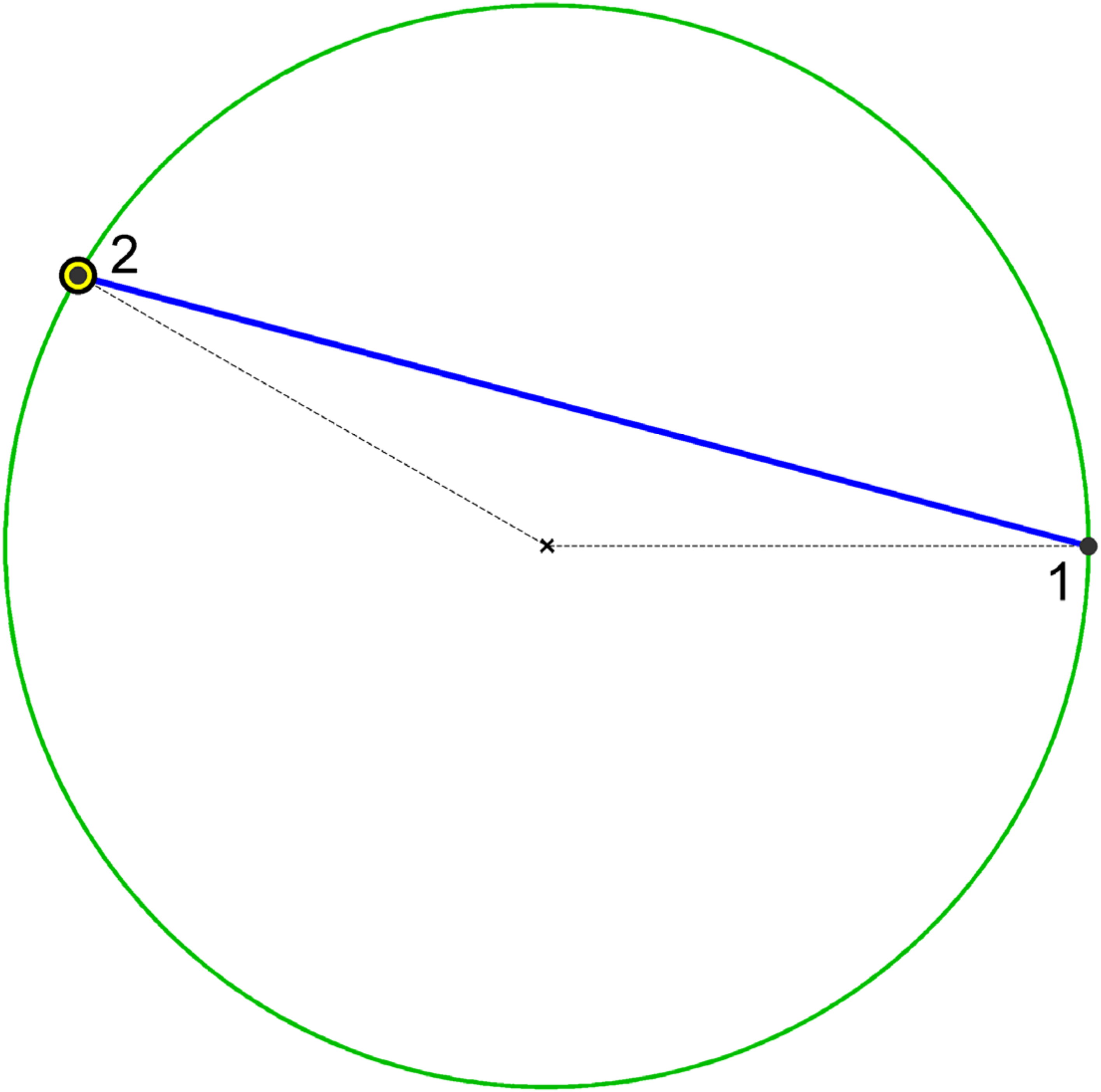
Mechanism (contd.)

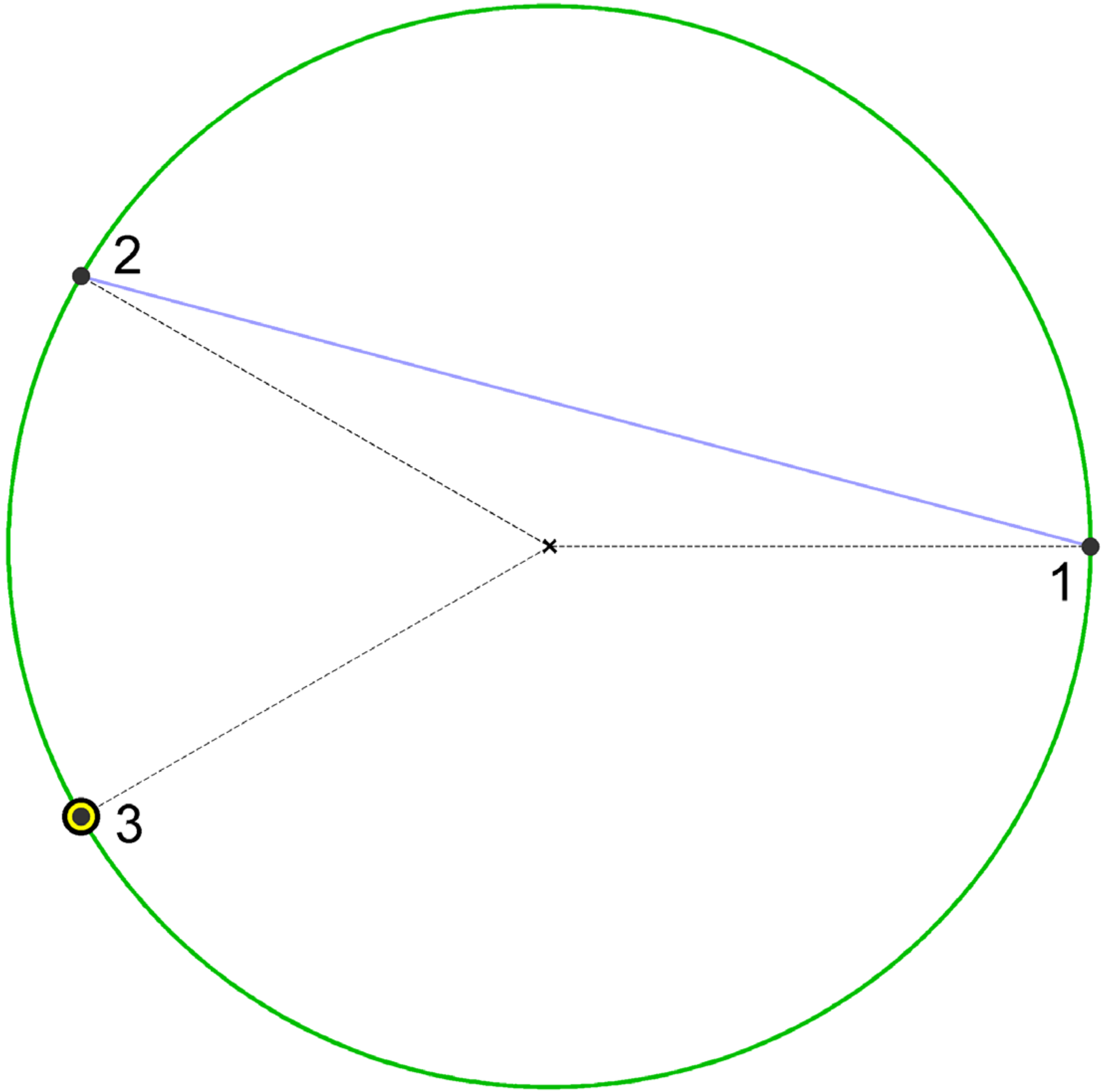
- New connections
  - New node t connects to m existing nodes s, s ? t, minimizing $s\theta_{st}$
  - That is, maximizing the product between popularity and similarity

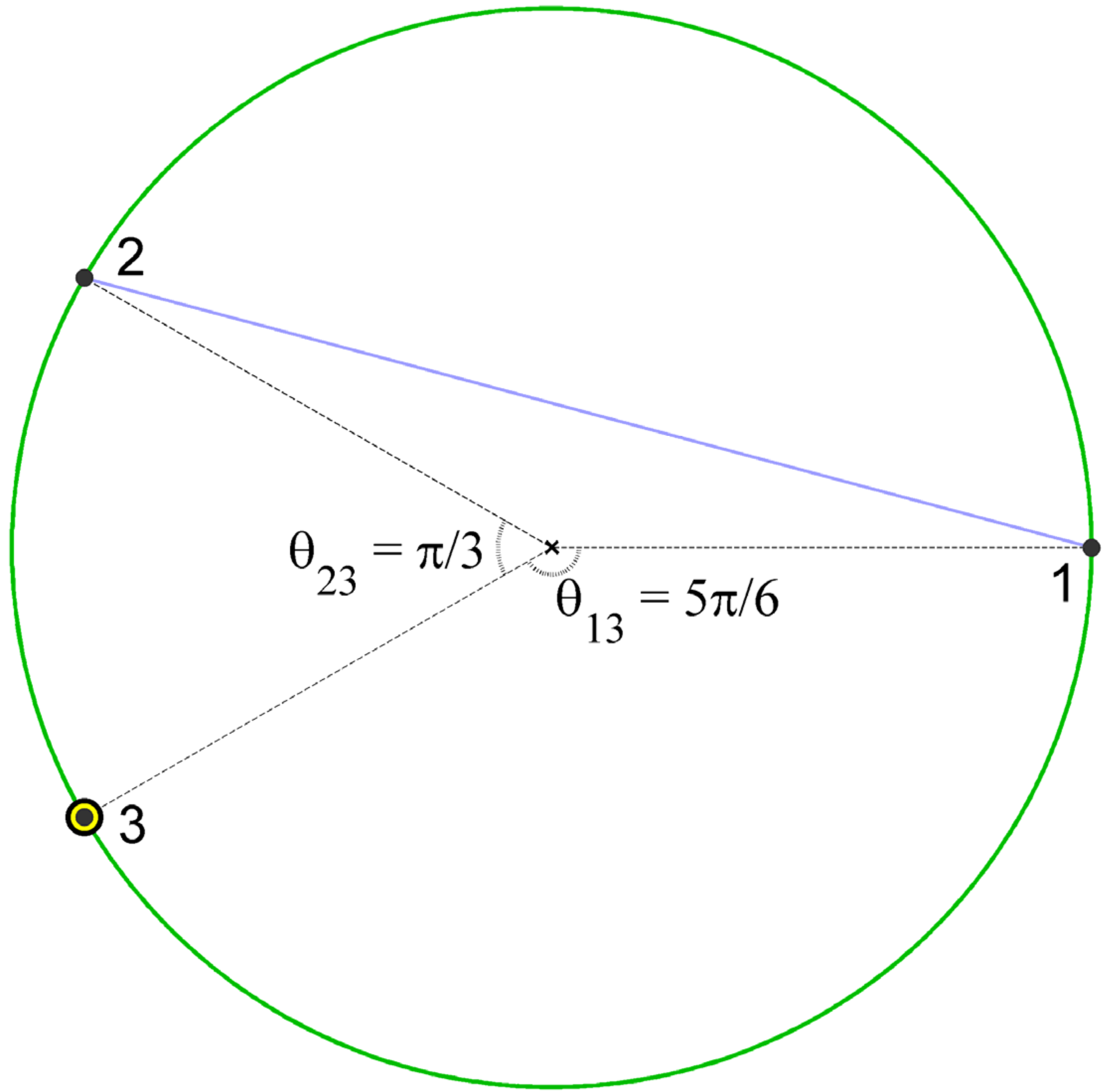


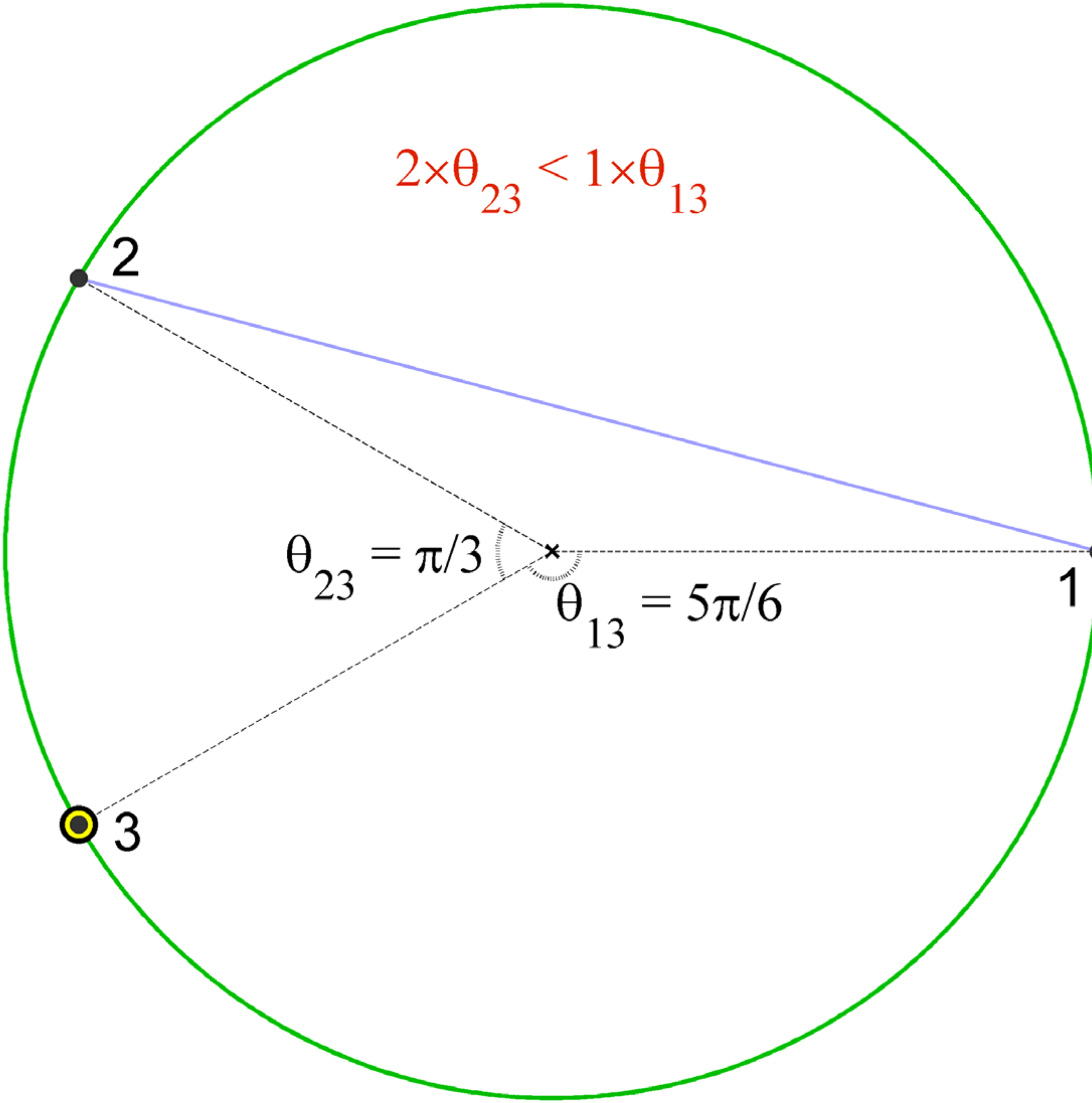

1


2
1

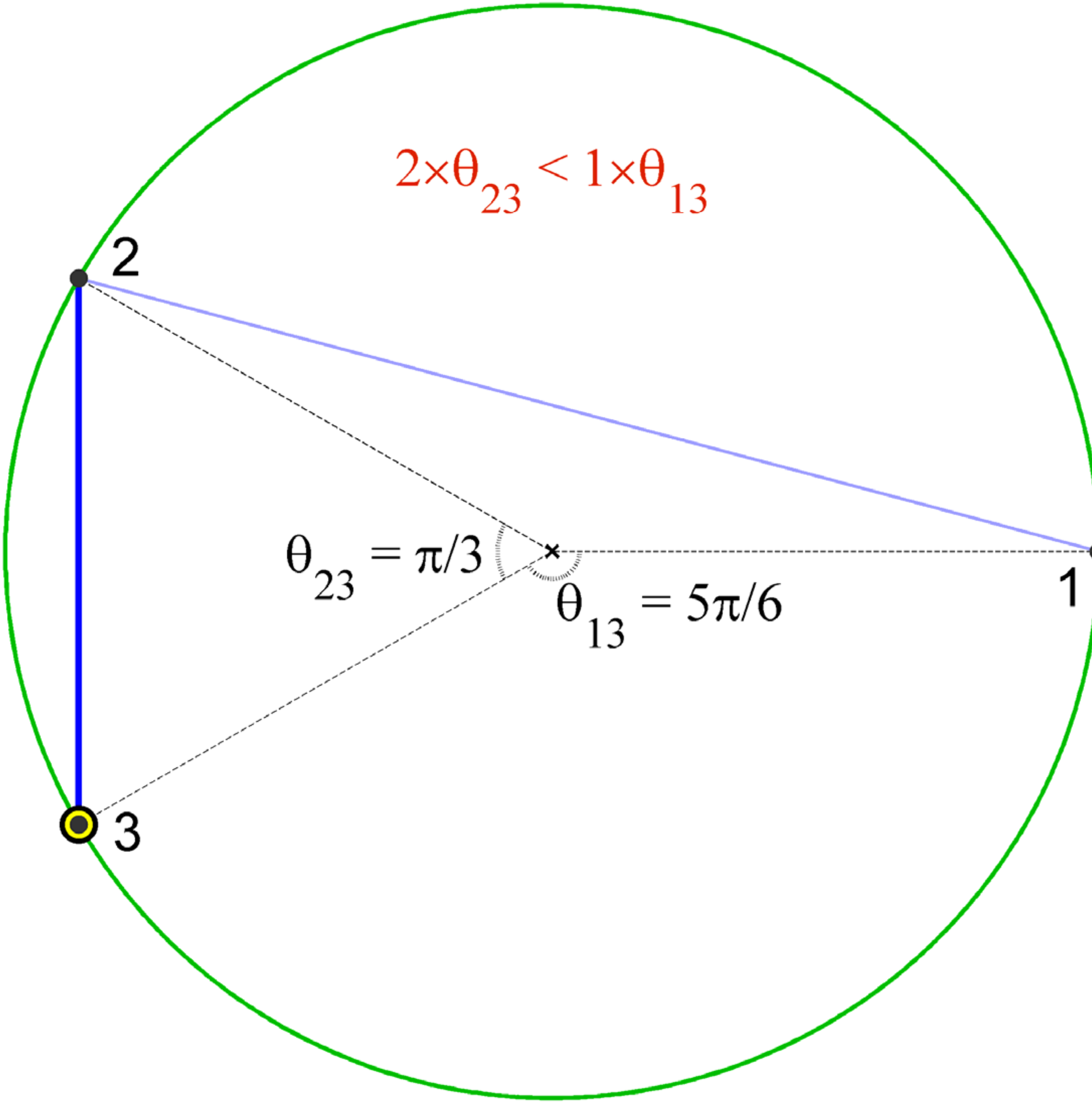

2
1

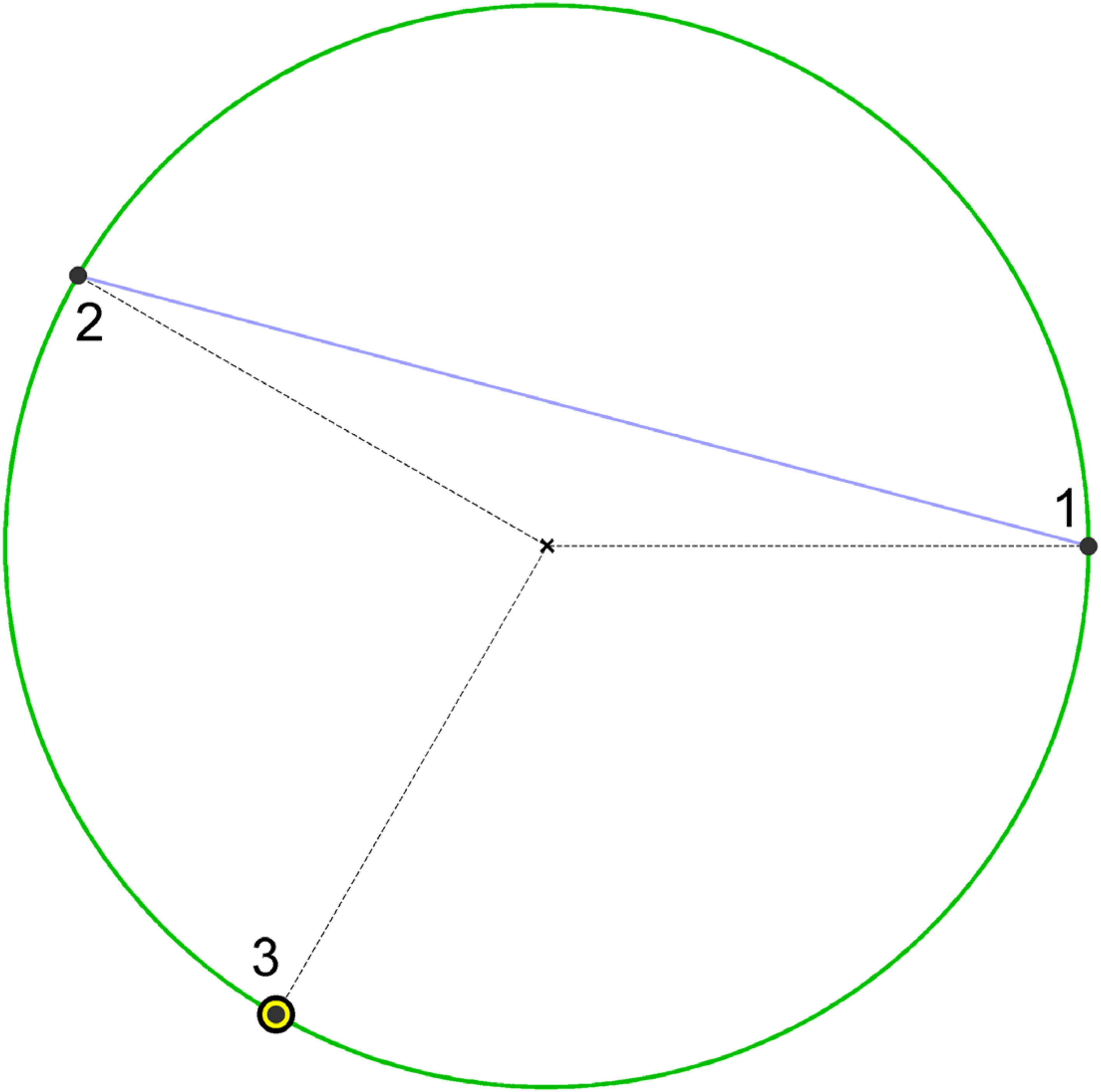

2
1
3

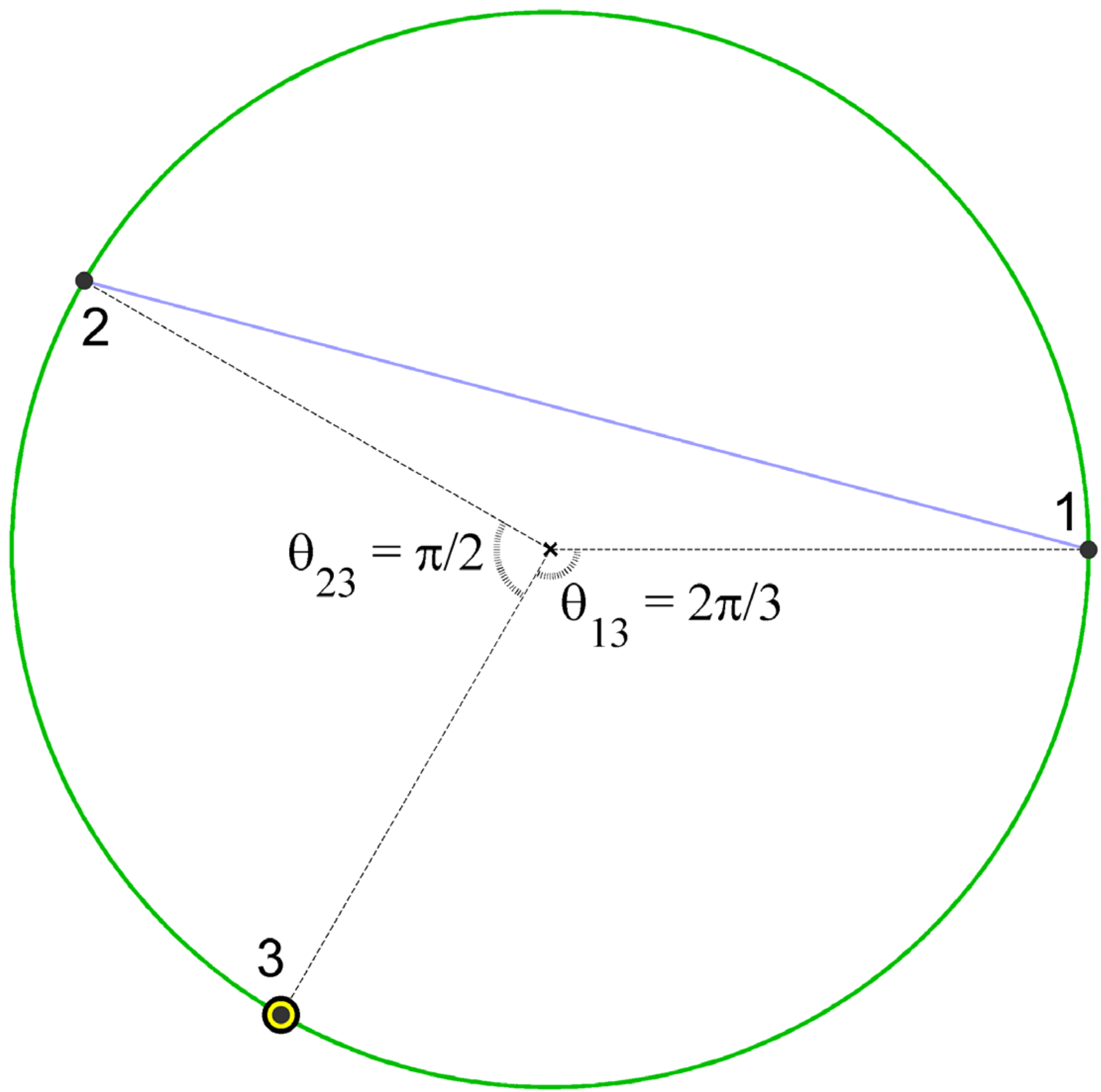

2
$\theta_{23} = \pi/3$
$\theta_{13} = 5\pi/6$
1
3

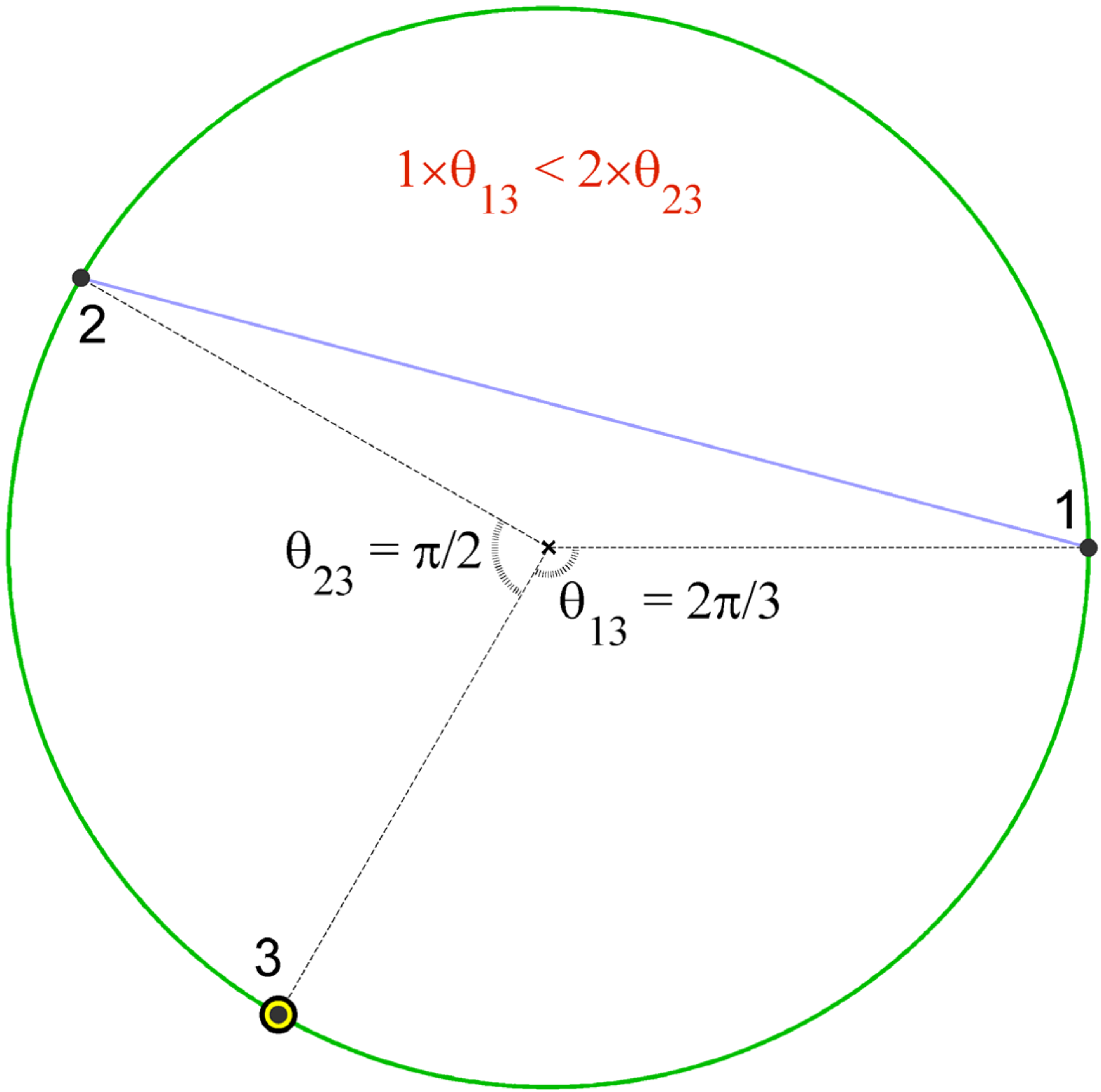


$2\times\theta_{23} < 1\times\theta_{13}$
2
$\theta_{23} = \pi/3$
$\theta_{13} = 5\pi/6$
1
3

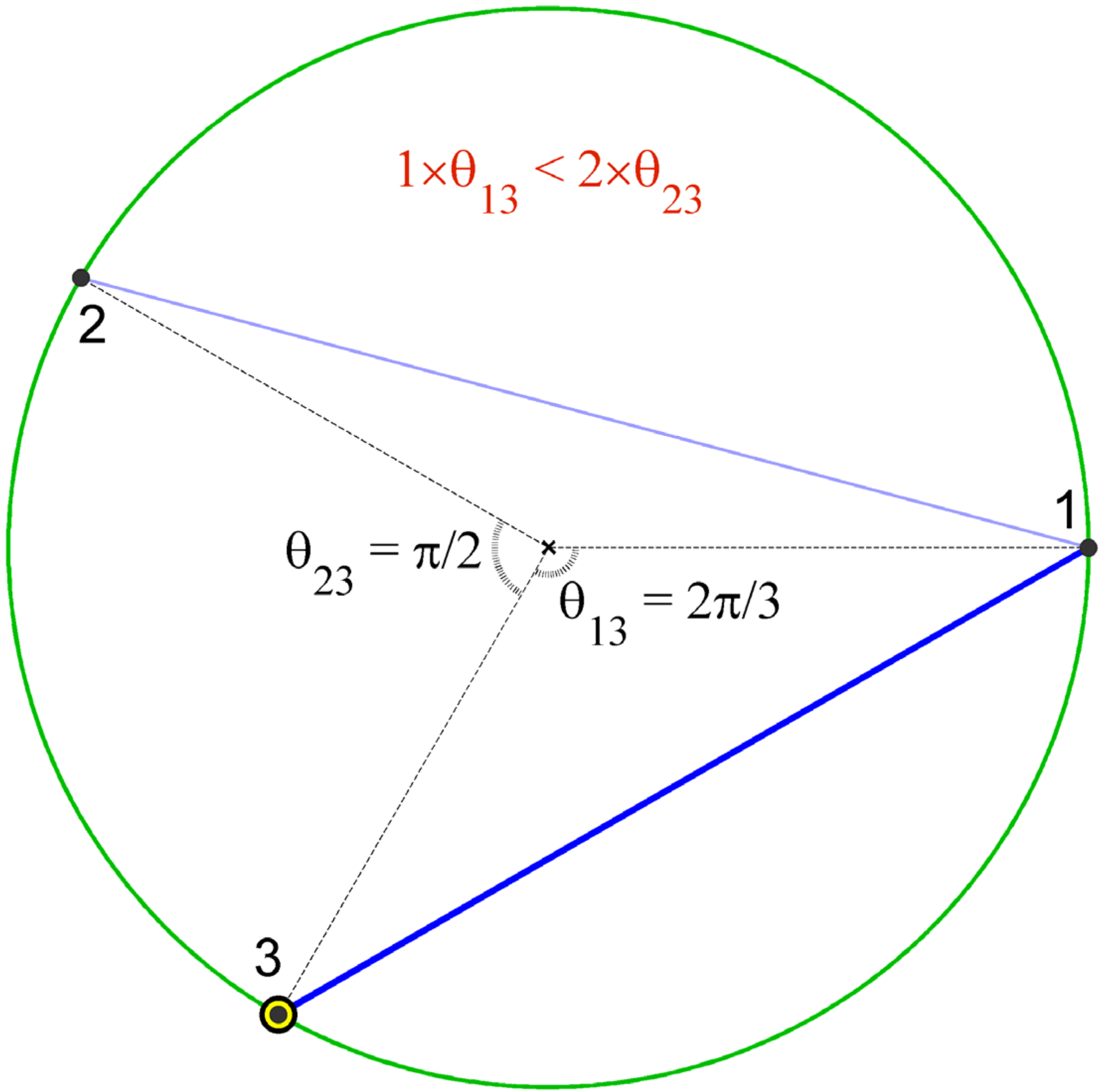

$2\times\theta_{23} < 1\times\theta_{13}$
2
$\theta_{23} = \pi/3$
$\theta_{13} = 5\pi/6$
1
3


2
1
3

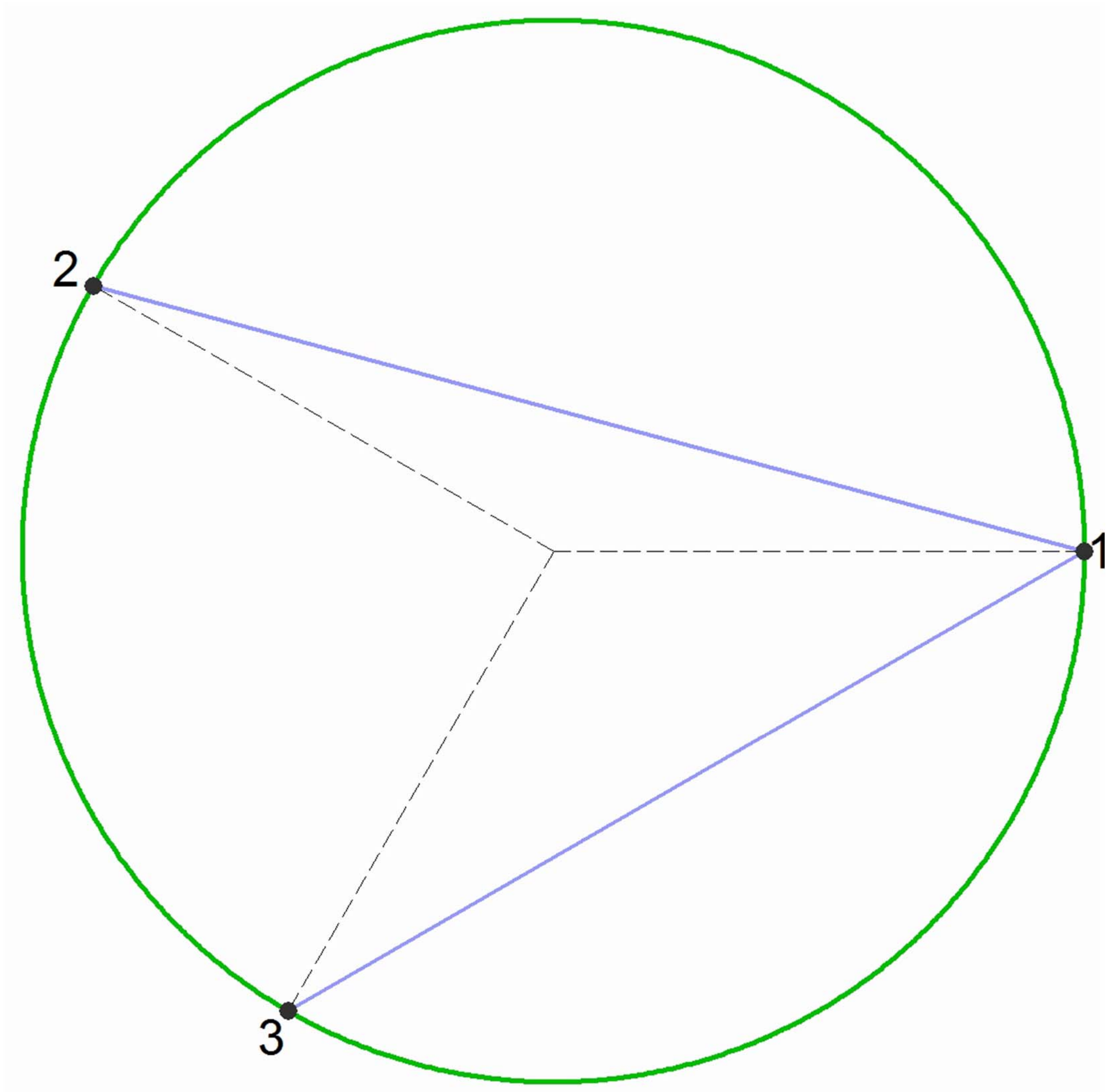

2
1
θ23 = π/2
θ13 = 2π/3
3



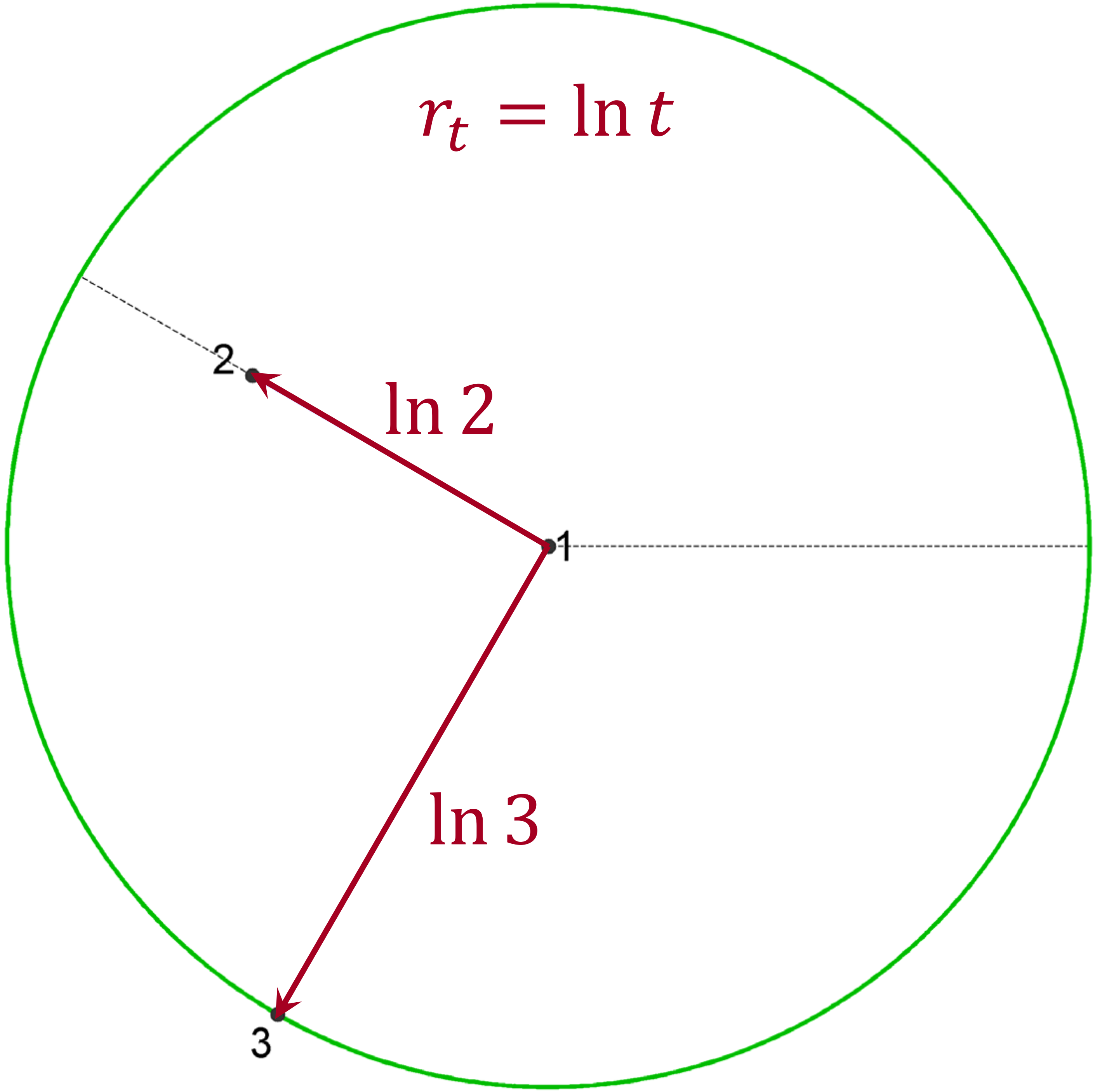
$1 \times \theta_{13} < 2 \times \theta_{23}$
2
1
$\theta_{23} = \pi/2$
$\theta_{13} = 2\pi/3$
3



$1\times\theta_{13} < 2\times\theta_{23}$
2
1
$\theta_{23} = \pi/2$
$\theta_{13} = 2\pi/3$
3



2
1
3


$r_t = \ln t$
2
$\ln 2$
1
$\ln 3$
3

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

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

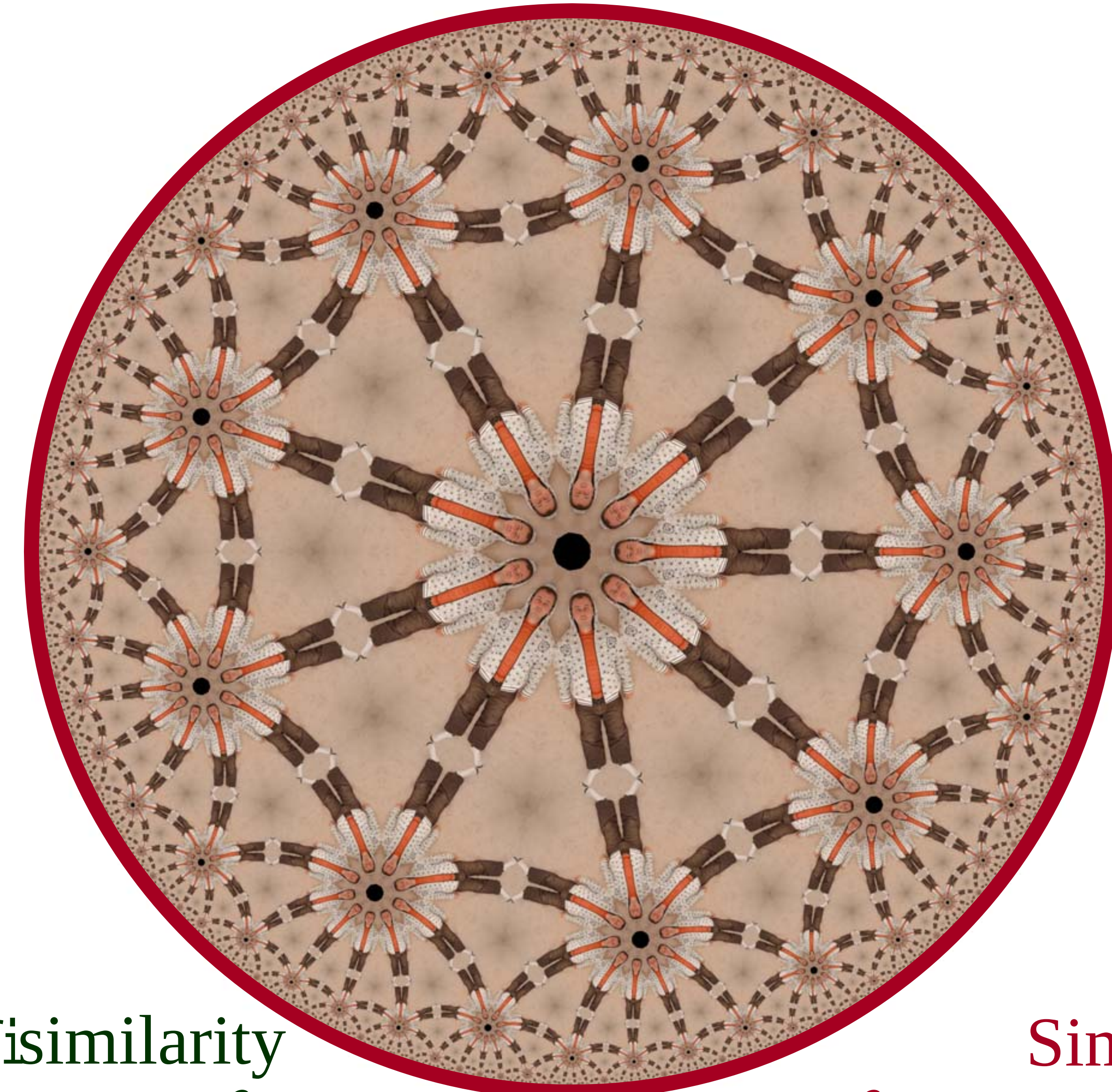
$$R_t = r_t - \ln \frac{2r_t}{\pi m}$$ — average degree is fixed to $2m$

$$R_t = r_t$$ — average degree grows logarithmically with *t* if ј ¬ 2

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

H^2

∂H^2

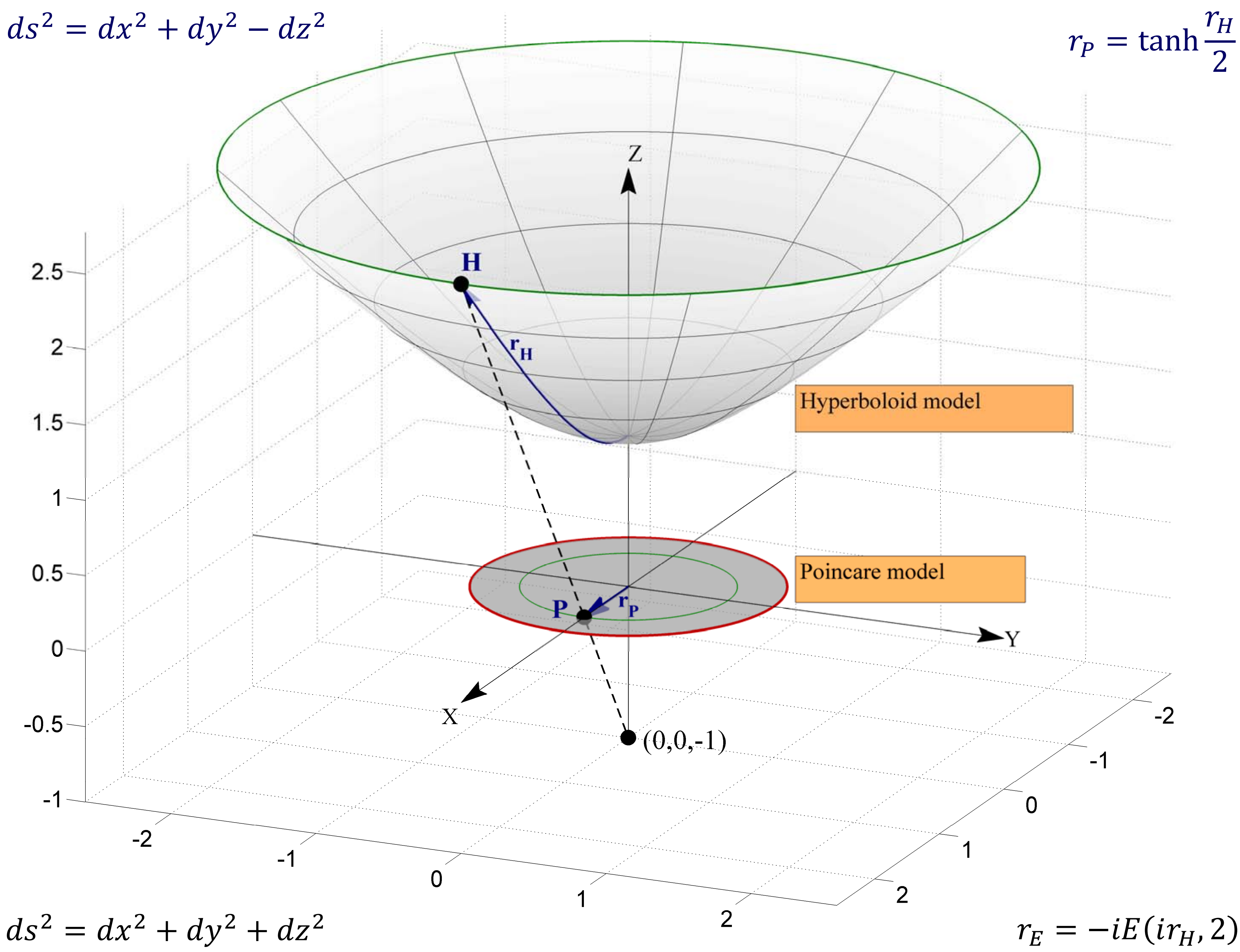


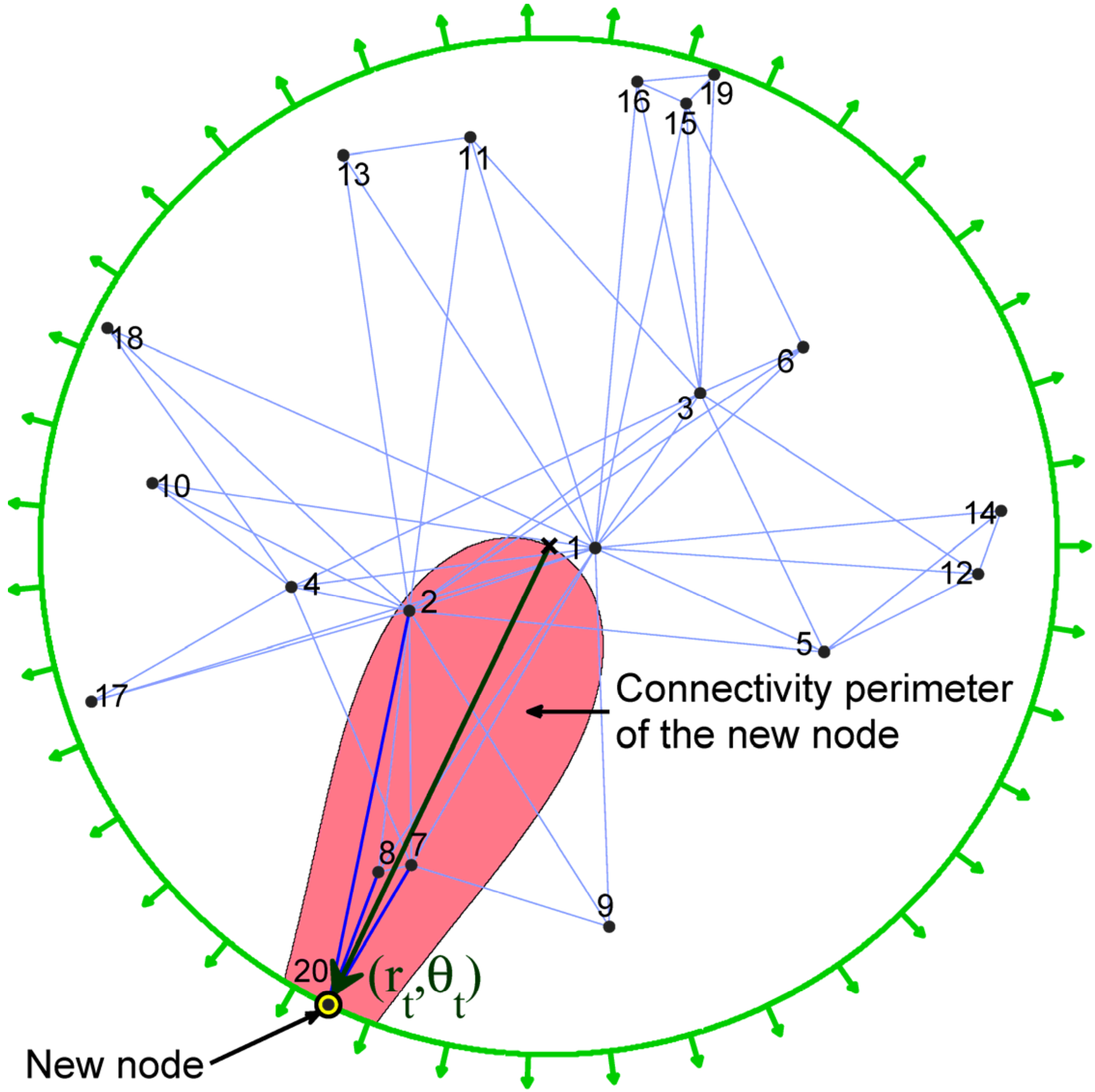
Popularityfisimilarity

$$x_{st} \approx r_s + r_t + \ln \sin \frac{\theta_{st}}{2}$$

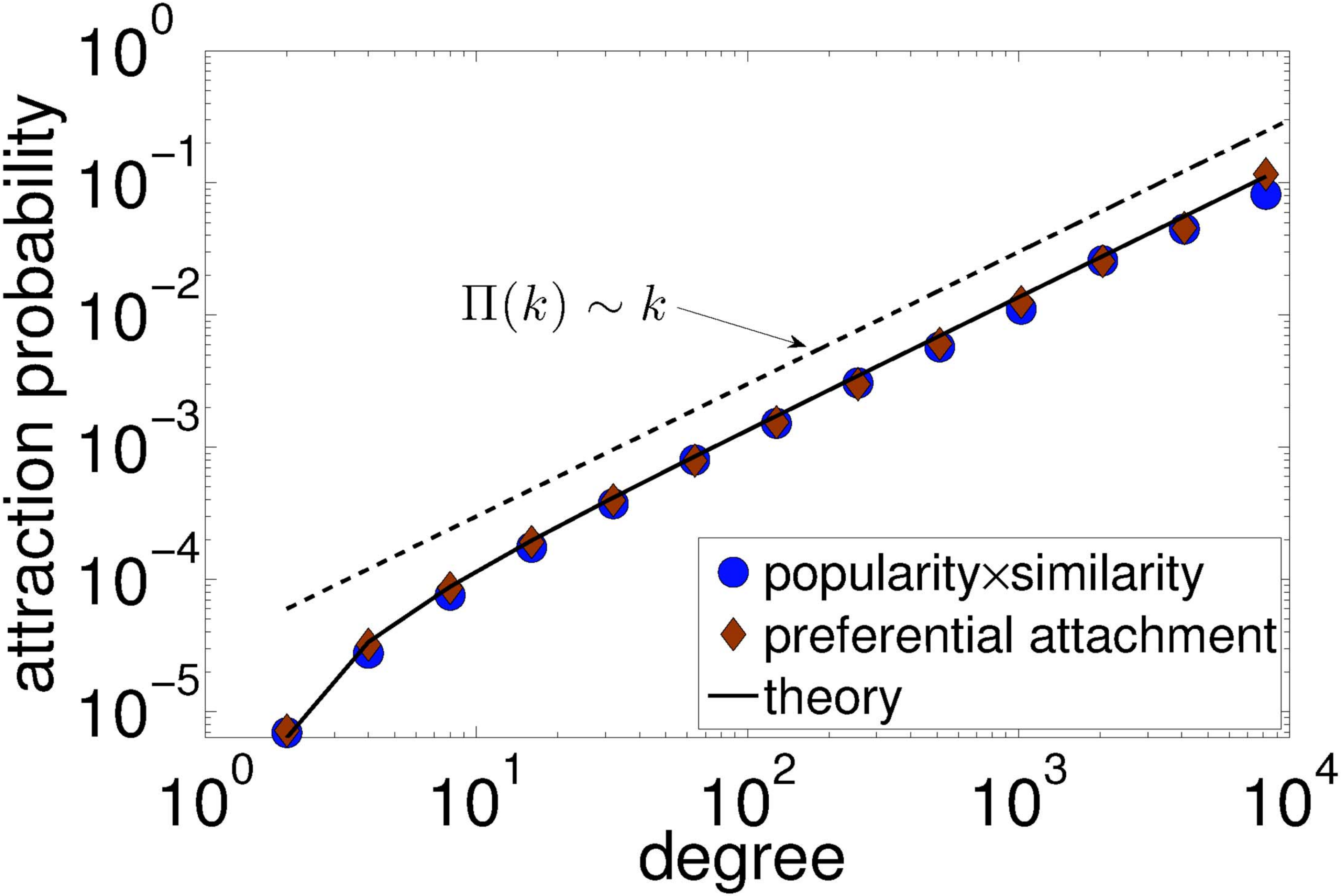
Similarity only

$$d_{st} = \sin \frac{\theta_{st}}{2} = \sqrt{(1 - \cos \theta_{st})/2}$$

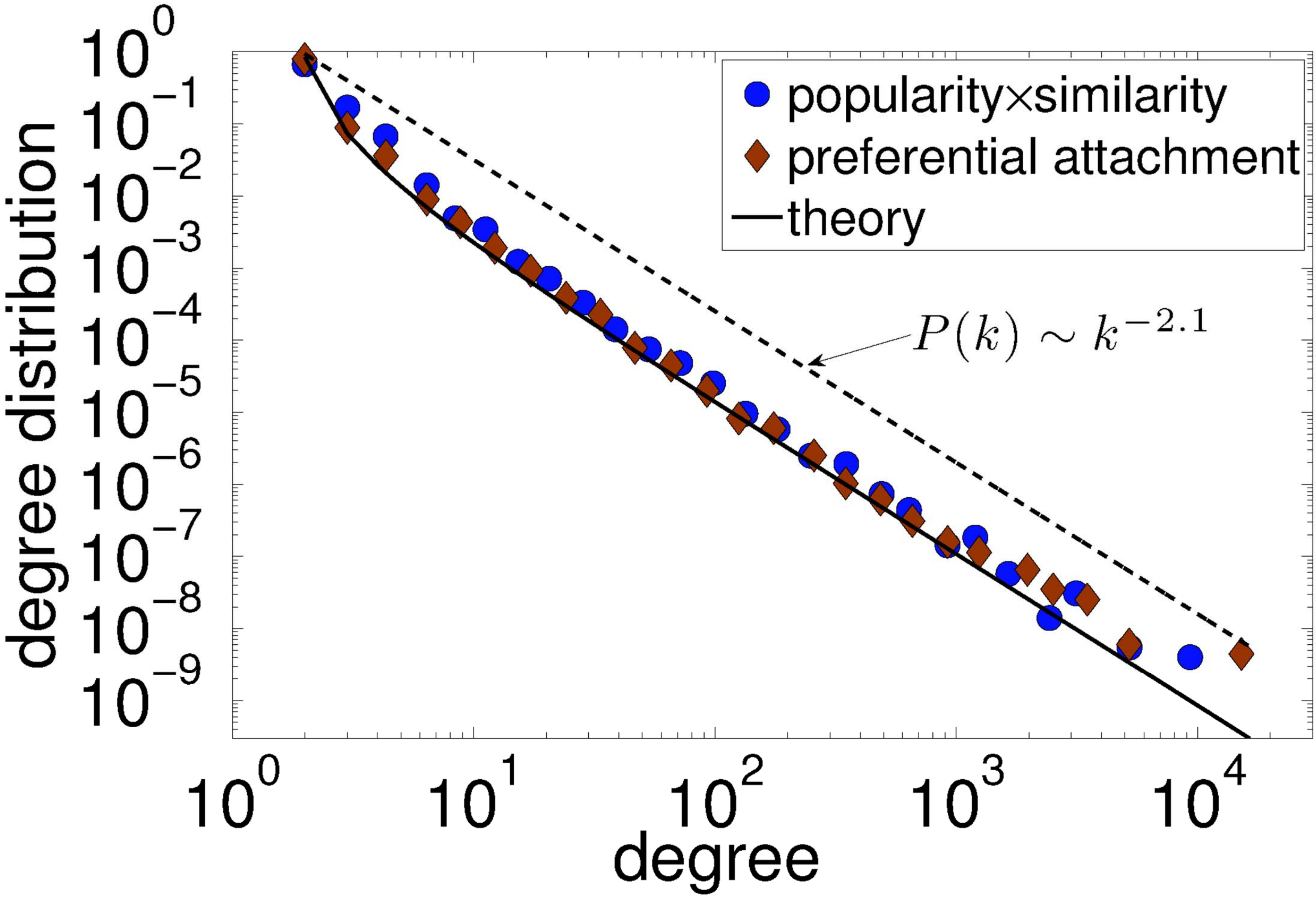

$ds^2 = dx^2 + dy^2 - dz^2$
$r_P = \tanh \frac{r_H}{2}$
Z
H
r_H
Hyperboloid model
Poincare model
P
r_P
Y
X
(0,0,-1)
$ds^2 = dx^2 + dy^2 + dz^2$
$r_E = -iE(ir_H, 2)$



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



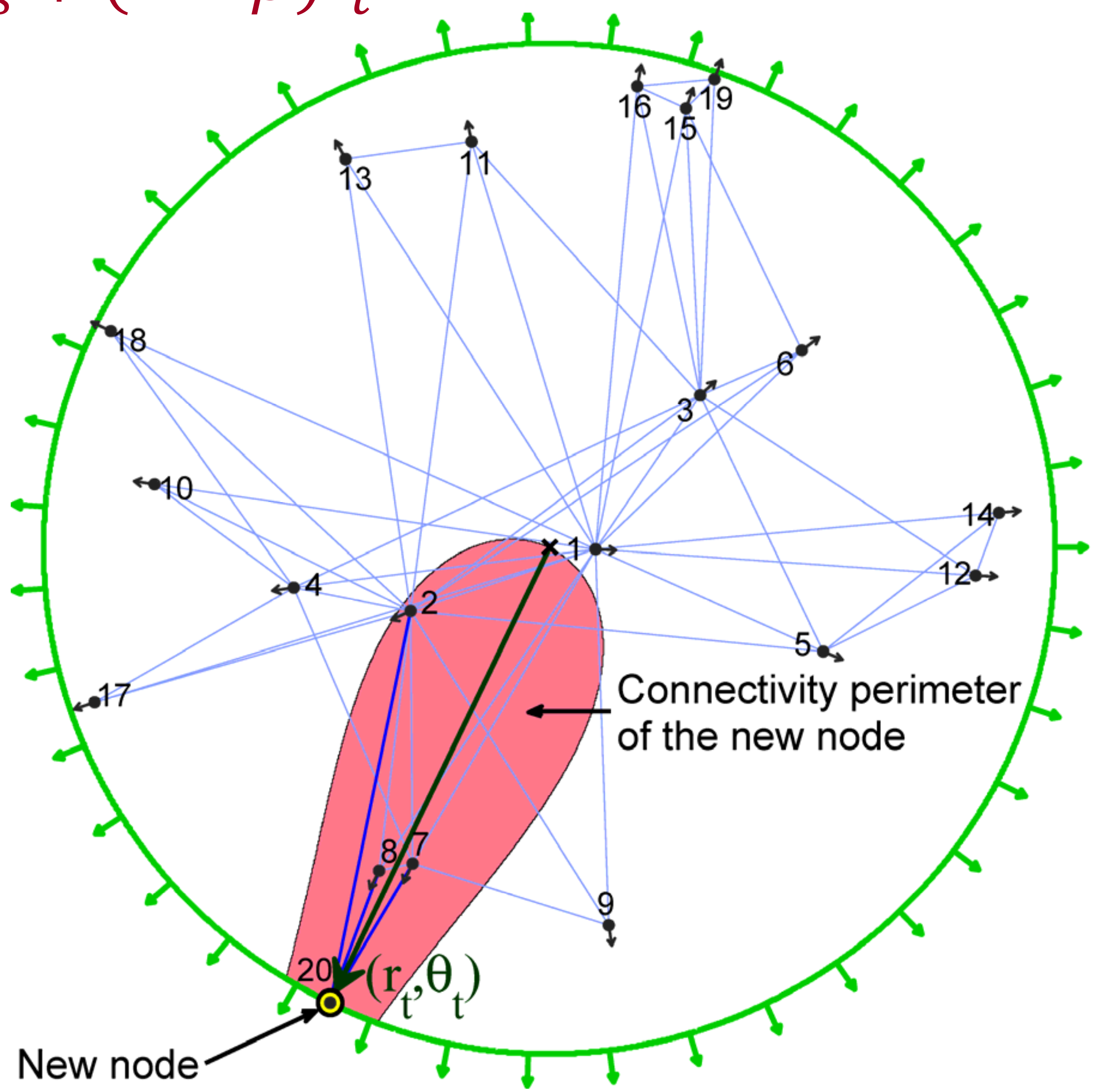
attraction probability
degree
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^0
10^1
10^2
10^3
10^4
$\Pi(k) \sim k$
popularity×similarity
preferential attachment
theory

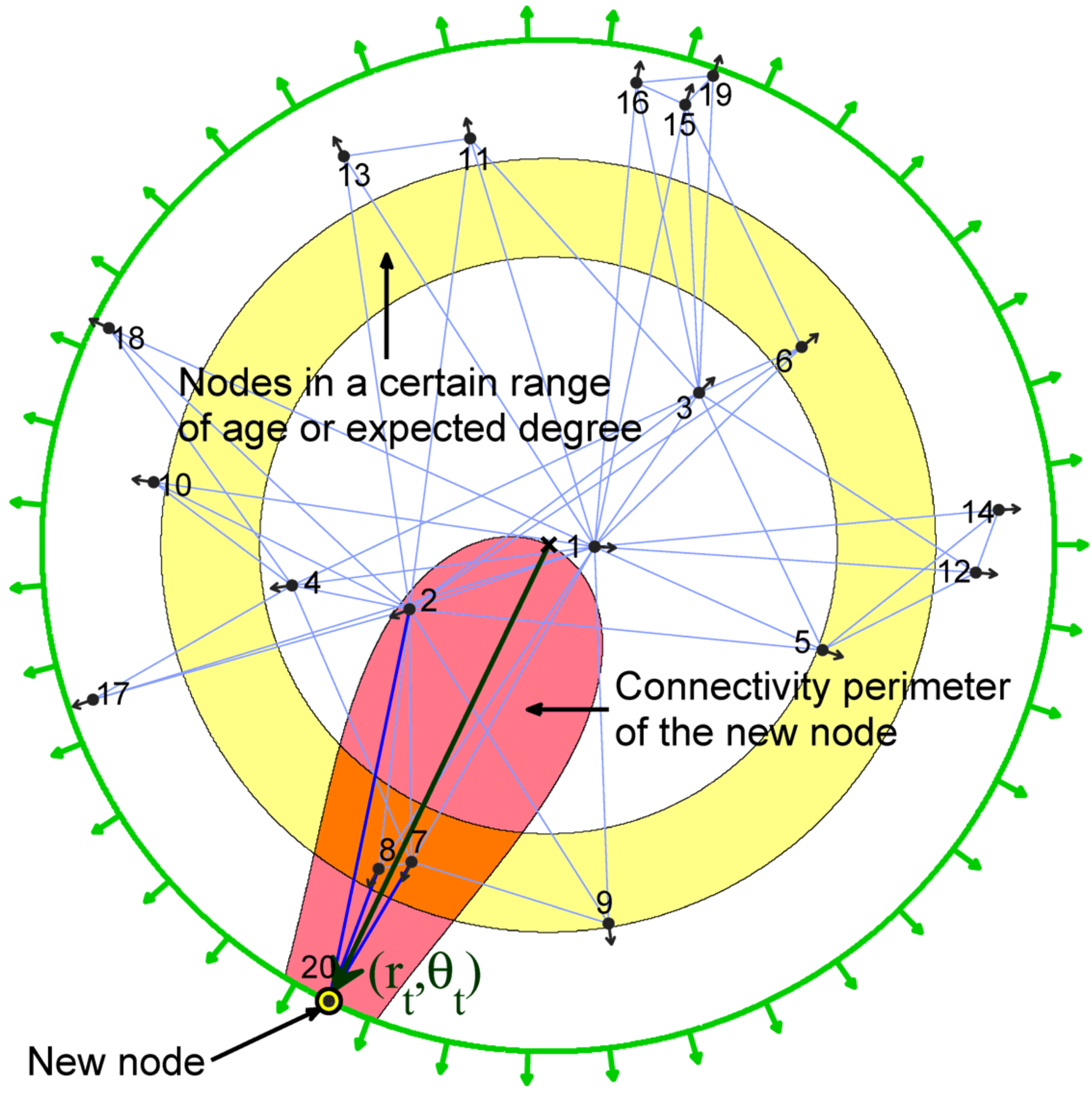


popularity×similarity
preferential attachment
theory
$P(k) \sim k^{-2.1}$
degree distribution
degree
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^0
10^1
10^2
10^3
10^4

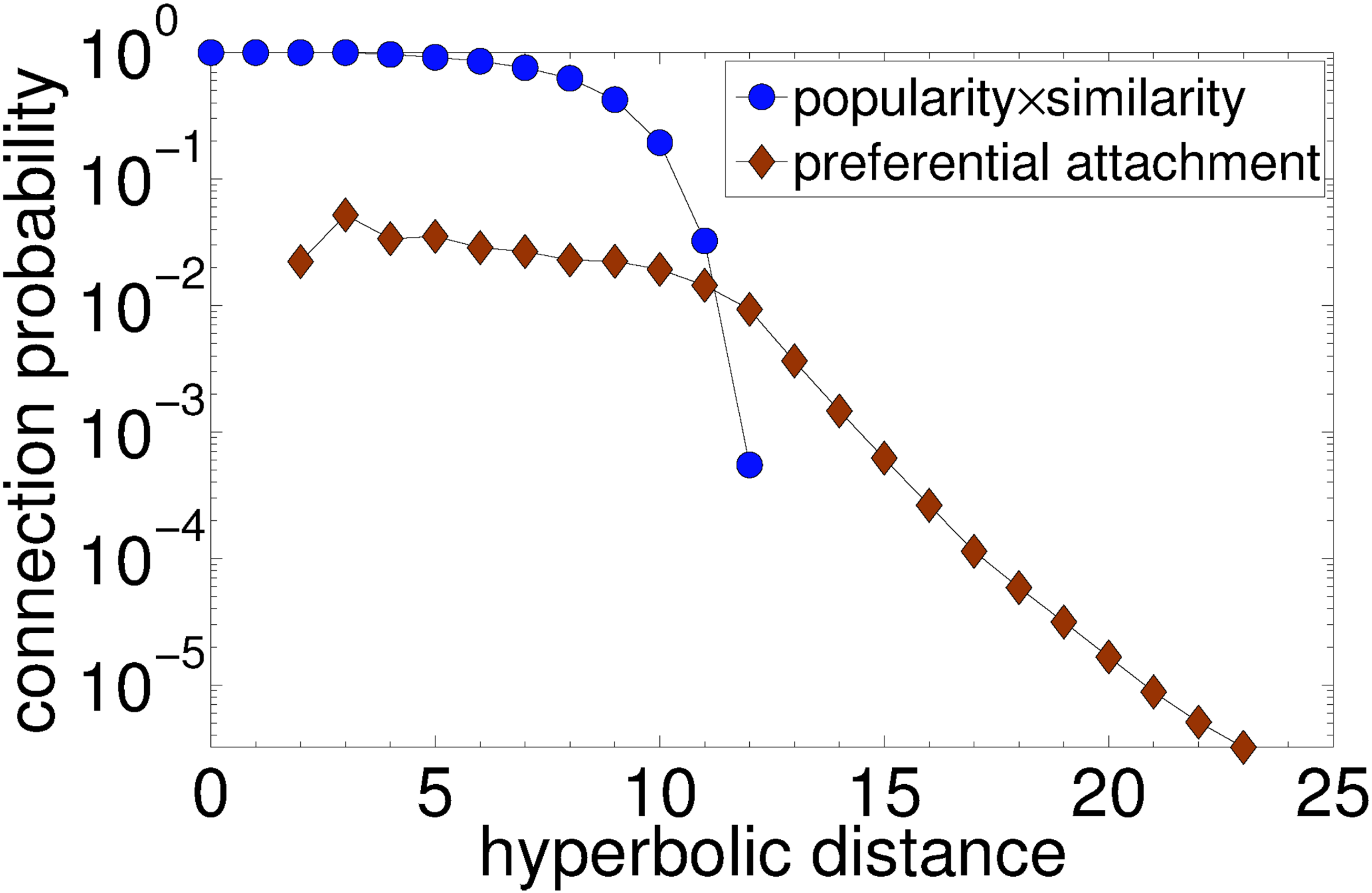
$$r_s(t) = \beta r_s + (1-\beta) r_t$$
$$s < t$$
$$0 < \beta \leq 1$$

$$\gamma = 1 + \frac{1}{\beta}$$
$$P(k) \sim k^{-\gamma}$$

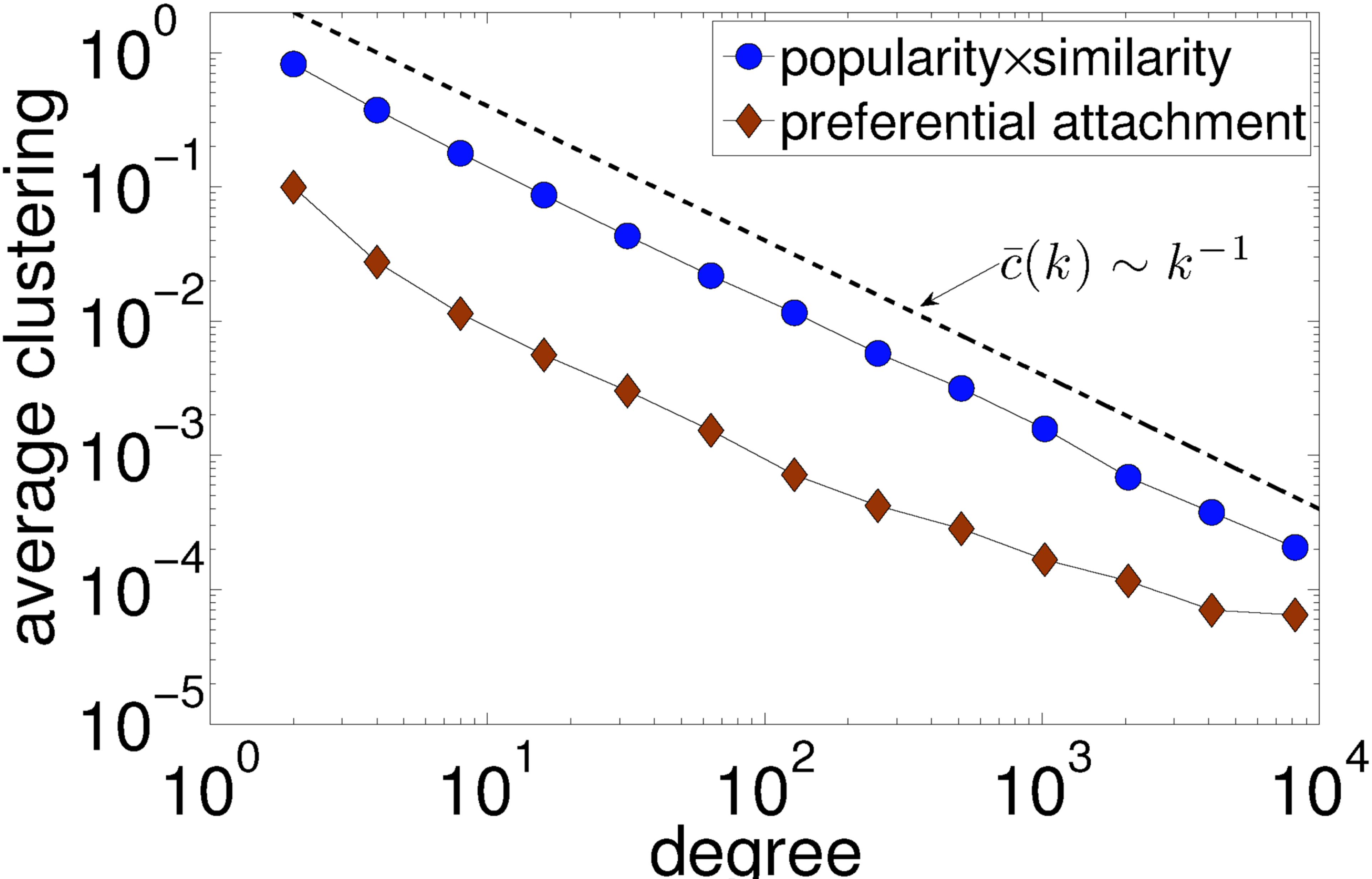




16
19
15
13
11
18
6
Nodes in a certain range
of age or expected degree
3
10
14
1
12
4
2
5
17
Connectivity perimeter
of the new node
8
7
9
20
(r_t, θ_t)
New node



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



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

Clustering

- Probability of new connections from *t* to *s* so far

$$p(x_{st}) = \Theta(R_t - x_{st})$$

- If we smoothen the threshold

$$p(x_{st}) = \frac{1}{1 + e^{\frac{x_{st} - R_t}{T}}} \xrightarrow{\mathrm{T} \to 0} \Theta(R_t - x_{st})$$

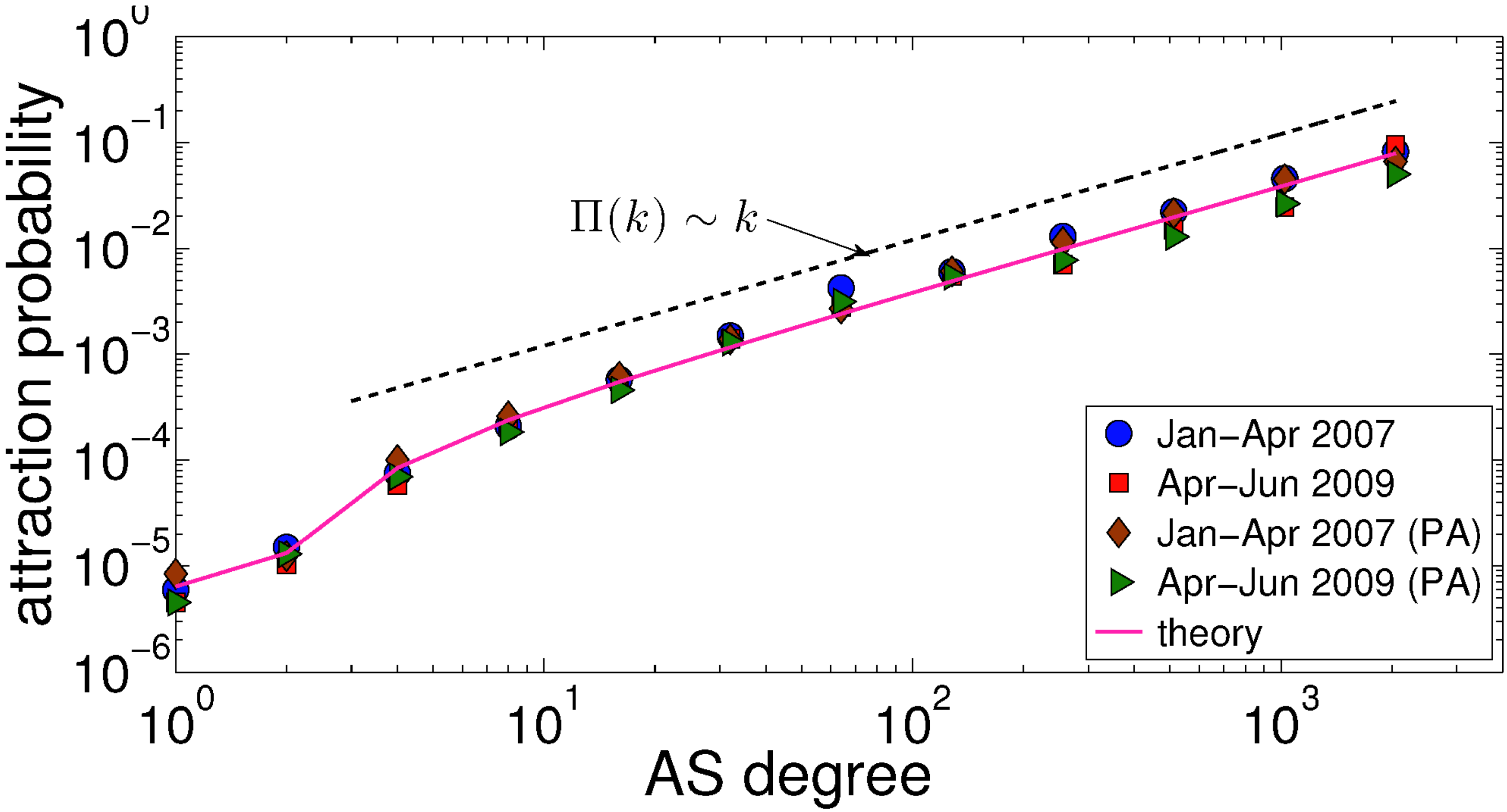
- Then average clustering linearly decreases with T from maximum at $T = 0$ to zero at $T = 1$
- Clustering is always zero at $T > 1$
- The model becomes identical to PA at *T* ´ –

Validation

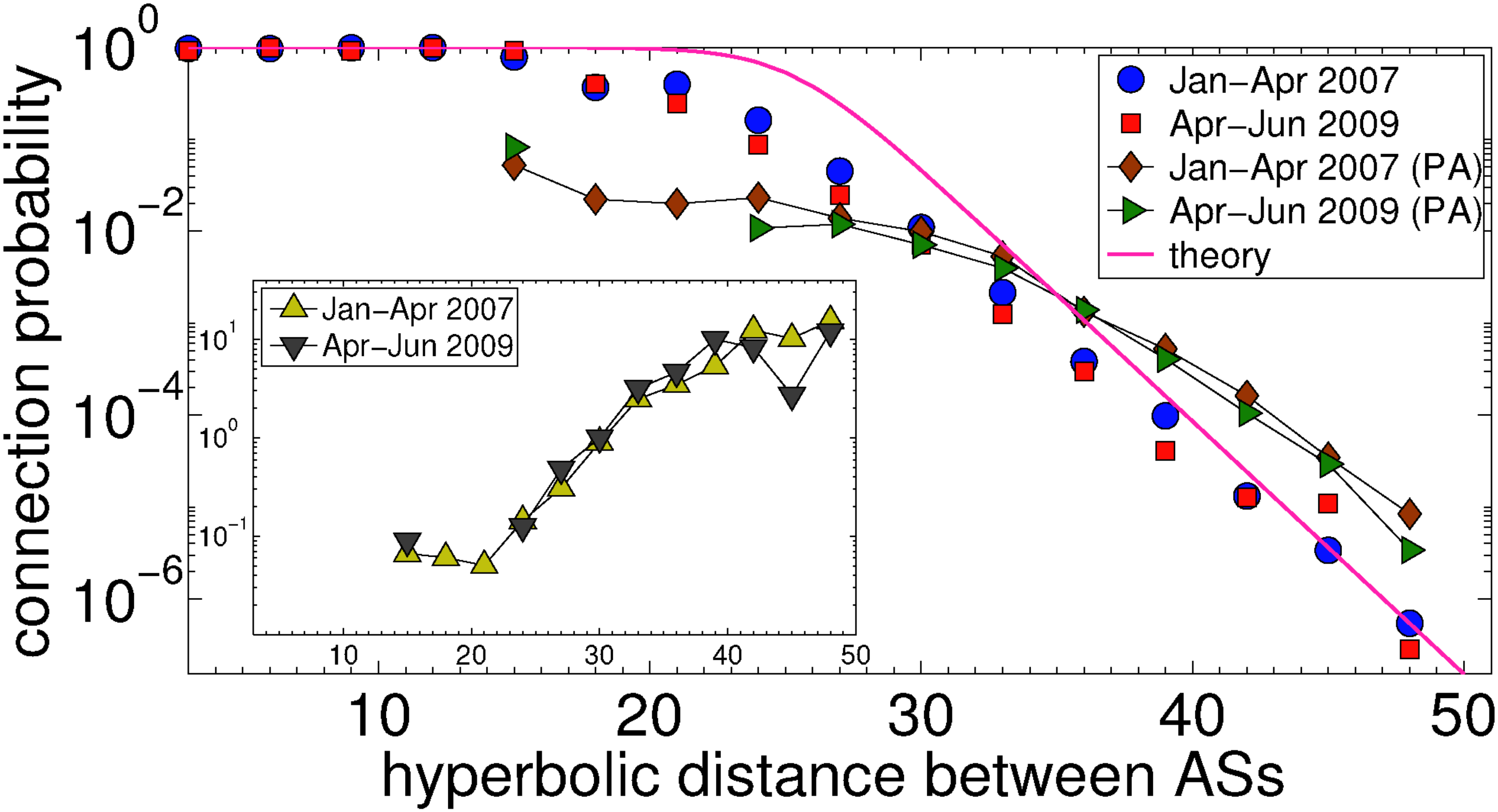
- Take a series of historical snapshots of a real network
- Infer angular/similarity coordinates for each node
- Test if the probability of new connections follows the model theoretical prediction

Learning similarity coordinates

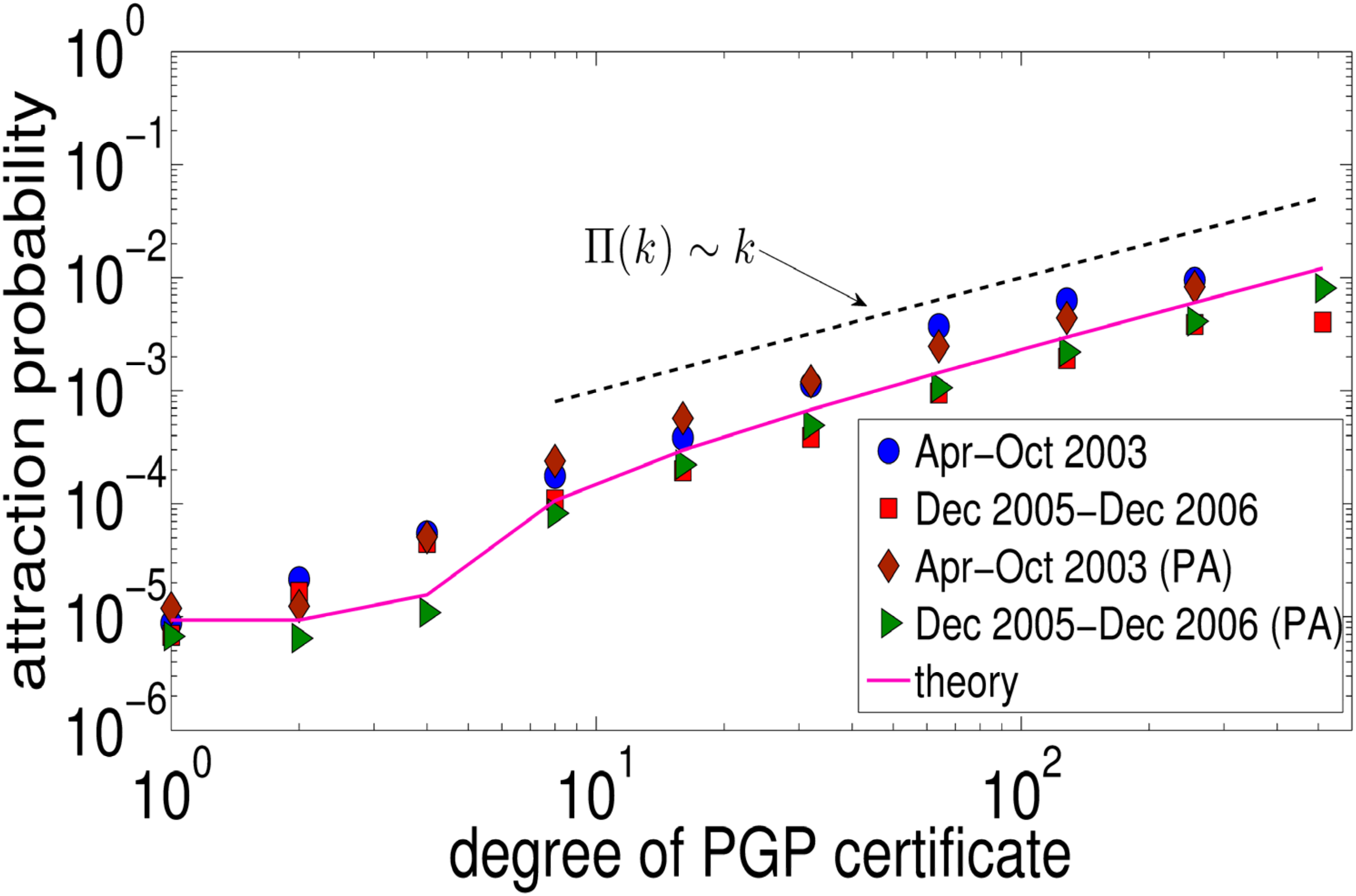
- Take a historical snapshot of a real network
- Apply a maximum-likelihood estimation method (e.g., MCMC) using the static hyperbolic model
- Metropolis-Hastings example
  - Assign random coordinates to all nodes
  - Compute current likelihood $L_c = \prod_{i<j} p(x_{ij})^{a_{ij}} [1 - p(x_{ij})]^{1-a_{ij}}$
  - Select a random node
  - Move it to a new random angular coordinate
  - Compute new likelihood L_n
  - If $L_n > L_c$, accept the move
  - If not, accept it with probability L_n / L_c
  - Repeat

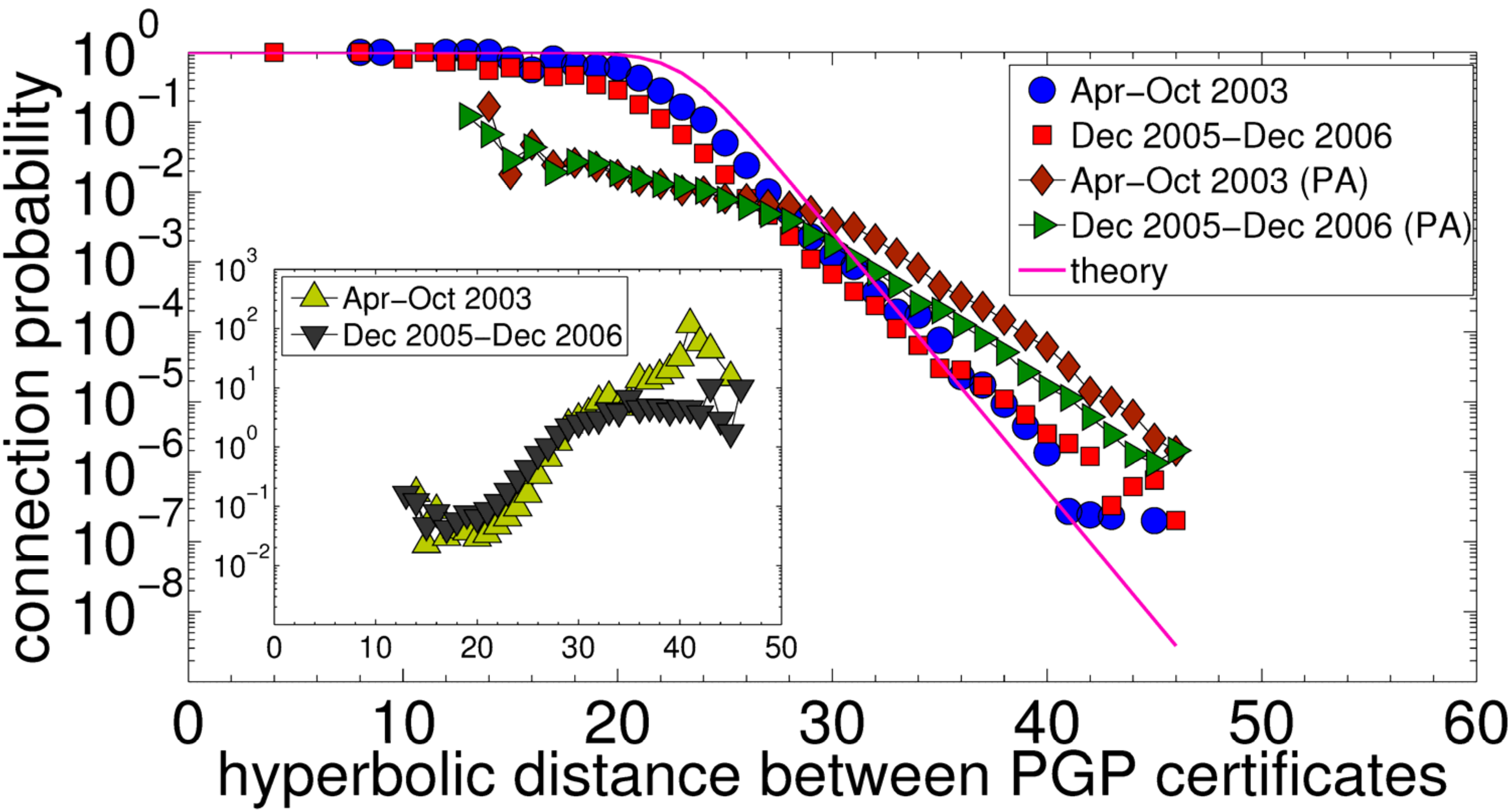


10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
attraction probability
$\Pi(k) \sim k$
Jan−Apr 2007
Apr−Jun 2009
Jan−Apr 2007 (PA)
Apr−Jun 2009 (PA)
theory
10^0
10^1
10^2
10^3
AS degree

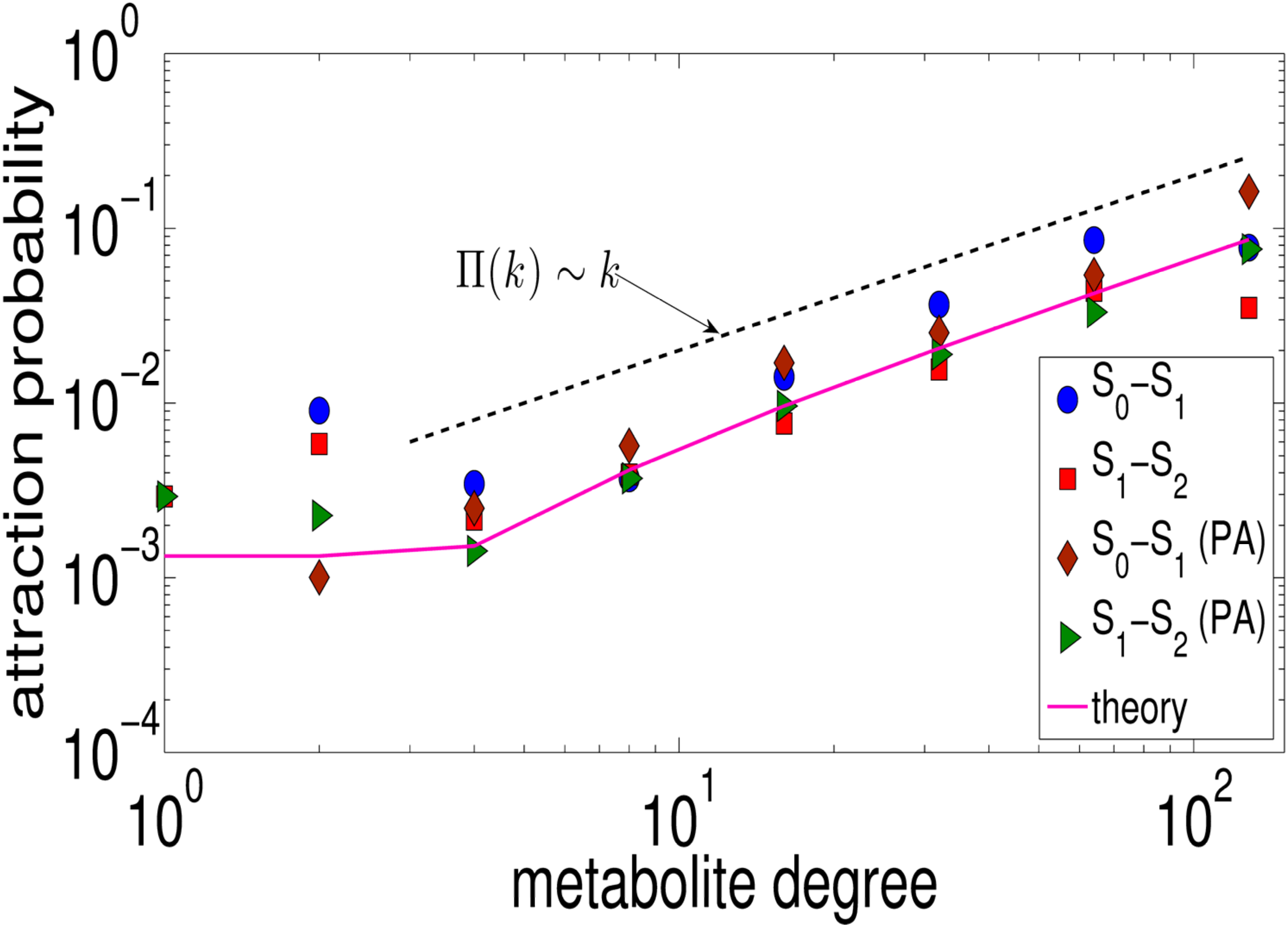


connection probability
hyperbolic distance between ASs
10^0
10^−2
10^−4
10^−6
10
20
30
40
50
Jan−Apr 2007
Apr−Jun 2009
Jan−Apr 2007 (PA)
Apr−Jun 2009 (PA)
theory
Jan−Apr 2007
Apr−Jun 2009
10^1
10^0
10^−1
10
20
30
40
50

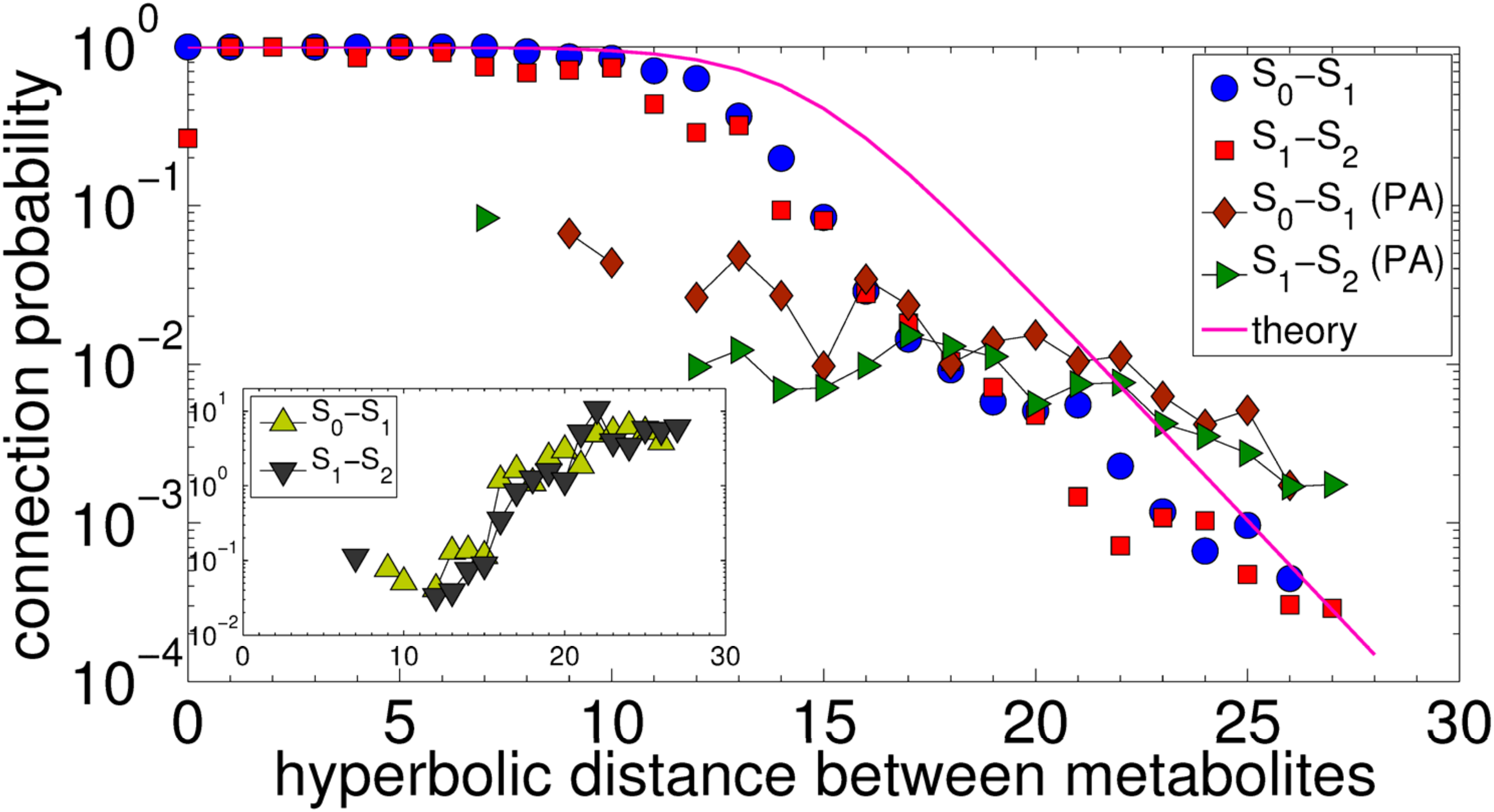

attraction probability
degree of PGP certificate
$\Pi(k) \sim k$
Apr−Oct 2003
Dec 2005−Dec 2006
Apr−Oct 2003 (PA)
Dec 2005−Dec 2006 (PA)
theory



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)

PSO compared to PA

- PA just ignores similarity, which leads to severe aberrations
  - Probability of similar connections is badly underestimated
  - Probability of dissimilar connections is badly overestimated
- If the connection probability is correctly estimated, then one immediate application is *link prediction*

Link prediction

- Suppose that some network has zero temperature
- Then one can predict links with 100% accuracy!
  - Because the connection probability is either 0 or 1

Non-zero temperature

- Link prediction is worse than 100%, but it must be still accurate since the connection probability is close to the step function
- No global intelligence is required
  - At zero temperature, new nodes connect to *exactly* the closest nodes
  - Non-zero temperature models reality where this hyperbolic proximity knowledge cannot be exact, and where it is mixed with errors and noise
- PA is an infinite-temperature regime with similarity forces reduced to nothing but noise

- M. Boguñá, F. Papadopoulos, and D. Krioukov,
  Sustaining the Internet with Hyperbolic Mapping,
  Nature Communications, v.1, 62, 2010

- D. Krioukov, F. Papadopoulos, A. Vahdat, and M. Boguñá,
  Hyperbolic Geometry of Complex Networks,
  Physical Review E, v.82, 036106, 2010,
  Physical Review E, v.80, 035101(R), 2009

- F. Papadopoulos, D. Krioukov, M. Boguñá, and A. Vahdat,
  Greedy Forwarding in (Dynamic) Scale-Free Networks Embedded in Hyperbolic Metric Spaces,
  INFOCOM 2010,
  SIGMETRICS MAMA 2009

- M. Boguñá, D. Krioukov, and kc claffy,
  Navigability of Complex Networks,
  Nature Physics, v.5, p.74-80, 2009