Inferring the origin of epidemic with dynamic message passing

Lenka Zdeborová
(IPhT, CNRS & CEA, Saclay, France)

with A. Lokhov, H. Ohta, M. Mezard.

Modeling the spread of an epidemic

SIR model: Kermack, McKendrick, 1927 - compartmental modeling. Since ~2001 hundreds of studies of SIR model on networks (Pastor-Satorras & Vespignani, May & Lloyd, Newman, etc. ....)

Assume we know the contact network: graph G(V,E), nodes i, edge ij. Each node i can be: S - susceptible, I - infected, R - recovered

The SIR dynamics

$S + I \xrightarrow{\lambda_{ij}} I + I$ $\quad$ λ_{ij} transmission probability

$I \xrightarrow{\mu_i} R$ $\quad$ μ_i recovery probability

With discrete time: $P(i : I \to R) = \mu_i$

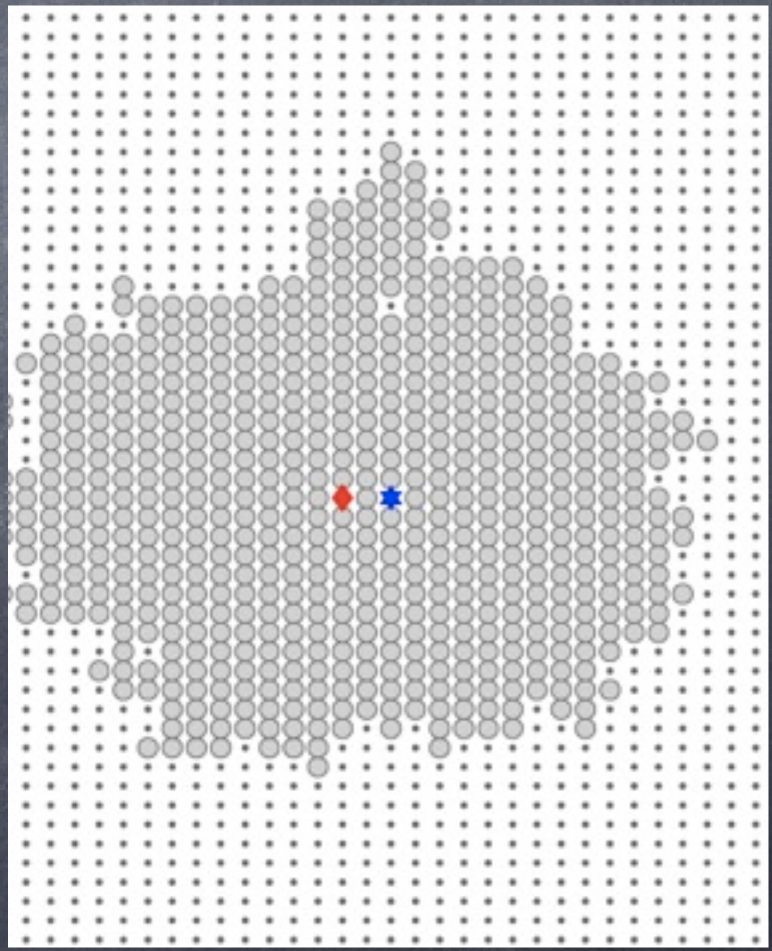
$$P(i : S \to I) = 1 - \prod_{j \in N(i)} (1 - \lambda_{ij} \delta_{I,q(j)})$$

Inference of epidemic origin

Time $t = 0$: node i infected, all others susceptible. Run SIR dynamics for t_0 steps, and observe the state of all nodes.

The statement of the problem

Given the contact network and the snapshot of states of all nodes (or of their fraction) infer which node was the origin.



from Prakash, Vrekeen, Faloustos, 2012

First ideas

- Jordan center minimizes Jordan centrality: $$J(i) = \max_{j \in \mathcal{G}} d(i,j)$$
- Distance center (center of mass) minimizes distance centrality: $$D(i) = \sum_{j \in \mathcal{G}} d(i,j)$$

Where $\mathcal{G}$ is the subgraph containing only the I and R nodes.

First ideas

- Jordan center minimizes Jordan centrality: $J(i) = \max_{j \in \mathcal{G}} d(i,j)$
- Distance center (center of mass) minimizes distance centrality: $D(i) = \sum_{j \in \mathcal{G}} d(i,j)$

Where $\mathcal{G}$ is the subgraph containing only the I and R nodes.

The probabilistically optimal solution

- Bayesian inference: P($i = i_0$| snapshot) = P (snapshot |$i = i_0$) P($i = i_0$) / P(snapshot)
- Maximum likelihood (ML): max of P($i = i_0$| snapshot)
- Main trouble: How to estimate P(snapshot |$i = i_0$)?

Existing works

- Problem introduced (as far as I know) by D. Shah, T. Zaman (2010). They introduced 'rumor centrality' that on regular trees is the ML solution.
- On general graphs (random, scale free, some benchmarks of real networks ...) rumor centrality performs basically the same as distance centrality (despite the claim in the abstract of the original paper, and in agreement with the simulations presented in the original paper).
- A number of consecutive works (relaxing some of the original assumptions, more general models, other kinds of centralities): Comin, Fontoura Costa 2011; Zhu, Ying, 2012, Prakash, Vreeken, Faloutsos 2012, Fioriti, Chinnici, 2012, Pinto, Thiran, Vetterli 2012, Dong, Zhang, 2013, ...

Our work

- Motivation: I like to develop algorithm that approach better the optimal maximum likelihood.
- The trouble was how to estimate P(snapshot $|i = i_0$)?
- "Mean field approximation"

$$P(\vec{q} = \vec{q}^{\,0}\,|i = i_0) = \prod_{i \in G} P(q_i = q_i^0 | i = i_0)$$

- How to compute the probability (over the runs of the dynamics, with a fixed initial condition and fixed network) that a given node is in a given state? Direct simulation is a possibility but a very heavy one.

Message passing

- Belief propagation (BP) - reinvented in many disciplines, studies for hundreds of problems, used in practice for many of them (cf. Cris Moore's talk).
- Belief propagation estimates exactly marginals of static probability distributions of tree-networks, and often also well on loopy networks.
- Same strategy for dynamical problems?
- Dynamical belief propagation (DBP) - i.e. BP where variables are the node-trajectories is studied, but algorithmically very heavy.
- DBP - simplifies into an easily tractable form is some special cases including the SIR.

Dynamical message passing for SIR

Related works:

- Volz (2008) and J. C. Miller (2010) - DMP equations averaged over graphs = exact "dynamical mean field" equations for SIR of tree-like random graphs.

- Karrer, Newman (2010) - non-averaged single instance equations presented for more general version of SIR - not tractable. Simplification for the canonical SIR only averaged over initial conditions and graphs.

- Ohta, Sasa (2009) - analogous equations for random field Ising model at zero temperature averaged over all regular random graphs.

Dynamical message passing for SIR

Related works:

- Volz (2008) and J. C. Miller (2010) - DMP equations averaged over graphs = exact "dynamical mean field" equations for SIR of tree-like random graphs.

- Karrer, Newman (2010) - non-averaged single instance equations presented for more general version of SIR - not tractable. Simplification for the canonical SIR only averaged over initial conditions and graphs.

- Ohta, Sasa (2009) - analogous equations for random field Ising model at zero temperature averaged over all regular random graphs.

Single instance and specified initial conditions necessary for our case (example from K-SAT: 20 years of works on replica symmetry breaking versus survey propagation)

Dynamical message passing for SIR

Auxiliary cavity graph (ACG): node j does not cause infection $\lambda_{ji} = 0$

$P_S^{i\to j}(t)$ prob. that node i is S at time t, in the ACG

$\theta^{i\to j}(t)$ prob. that i did not send infection to j up to t in the ACG

$\phi^{i\to j}(t)$ prob. that i is I and did not send infection to j up to time t in the ACG

Iterative equations:

$$P_S^{i\to j}(t+1) = P_S^i(0) \prod_{k\in\partial i\setminus j} \theta^{k\to i}(t+1)$$

$$\theta^{k\to i}(t+1) - \theta^{k\to i}(t) = -\lambda_{ki}\phi^{k\to i}(t)$$

$$\phi^{k\to i}(t) = (1-\lambda_{ki})(1-\mu_k)\phi^{k\to i}(t-1) - [P_S^{k\to i}(t) - P_S^{k\to i}(t-1)]$$

Initialization: $\theta^{k\to i}(0) = 1$ $\phi^{k\to i}(0) = \delta_{q_k(0),I}$

Dynamical message passing for SIR

Auxiliary cavity graph (ACG): node j does not cause infection $\lambda_{ji} = 0$

$P_S^{i\to j}(t)$ prob. that node i is S at time t, in the ACG

$\theta^{i\to j}(t)$ prob. that i did not send infection to j up to t in the ACG

$\phi^{i\to j}(t)$ prob. that i is I and did not send infection to j up to time t in the ACG

Iterative equations:

the non-trivial part

$$P_S^{i\to j}(t+1) = P_S^i(0) \prod_{k\in\partial i\backslash j} \theta^{k\to i}(t+1)$$

$$\theta^{k\to i}(t+1) - \theta^{k\to i}(t) = -\lambda_{ki}\phi^{k\to i}(t)$$

$$\phi^{k\to i}(t) = (1-\lambda_{ki})(1-\mu_k)\phi^{k\to i}(t-1) - [P_S^{k\to i}(t) - P_S^{k\to i}(t-1)]$$

Initialization: $\theta^{k\to i}(0) = 1$ $\phi^{k\to i}(0) = \delta_{q_k(0),I}$

Dynamical message passing for SIR

Finally compute:

$$P_S^i(t+1) = P_S^i(0) \prod_{k \in \partial i} \theta^{k \to i}(t+1)$$

$$P_R^i(t+1) = P_R^i(t) + \mu_i P_I^i(t)$$

$$P_I^i(t+1) = 1 - P_S^i(t+1) - P_R^i(t+1)$$

Repeat for all possible origins i, for each compute

$$E(i) = -\sum_j \log P^j_{q_j^0}(t_0)$$

Rank nodes according to E(i). To infer the age of the epidemic, minimize E(i) over t_0 .

Remarks about DMP

- Solving DMP for a given initial condition is as easy as running a single simulation of SIR.
- BP is iterated till convergence, whereas in DMP the iteration time corresponds to the real time.
- Works for arbitrary initial condition and even for networks changing in time (transmission probability can be arbitrarily time-dependent)
- Limitations: Contact network needs to be known. Corrections caused by loops hard to control. If probability of recovery also depends on neighbors simple exact equations on trees are not known (yet)!

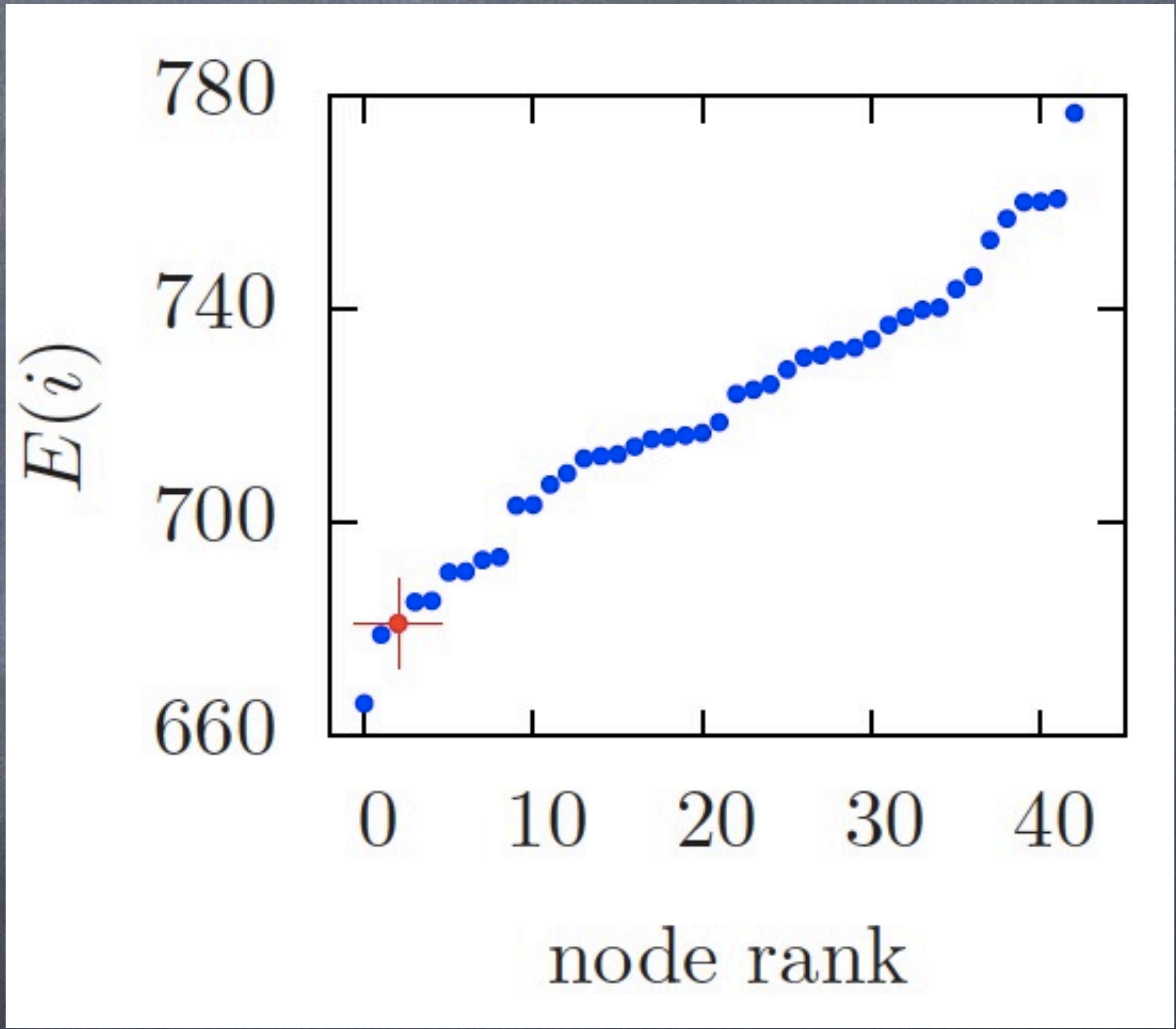
Results

random 4-regular graph, N=1000,

$\mu = 1, \lambda = 0.6$

$t_0 = 8$

242 nodes observed infected or recovered.



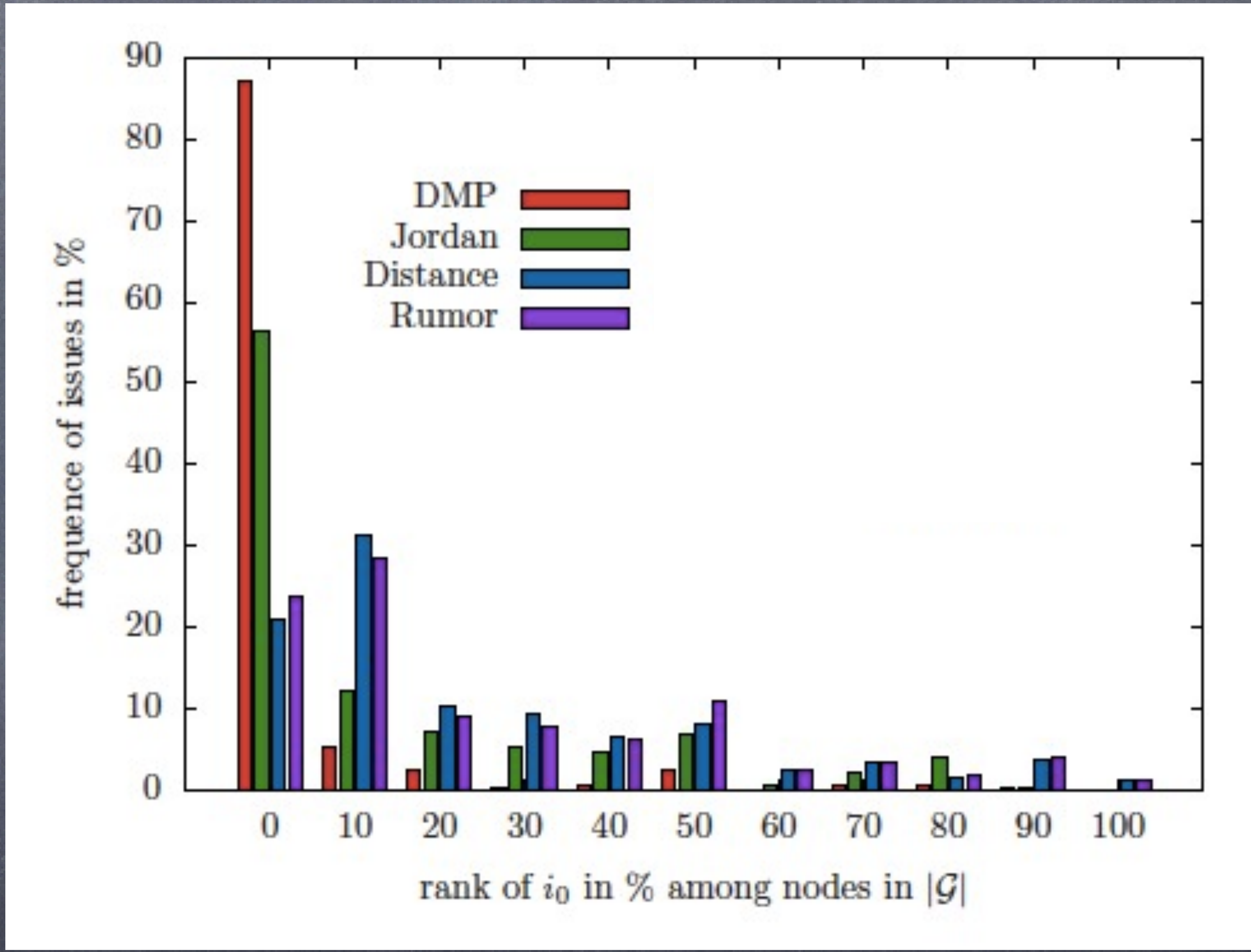
Results

random 4-regular graph,
N=1000,

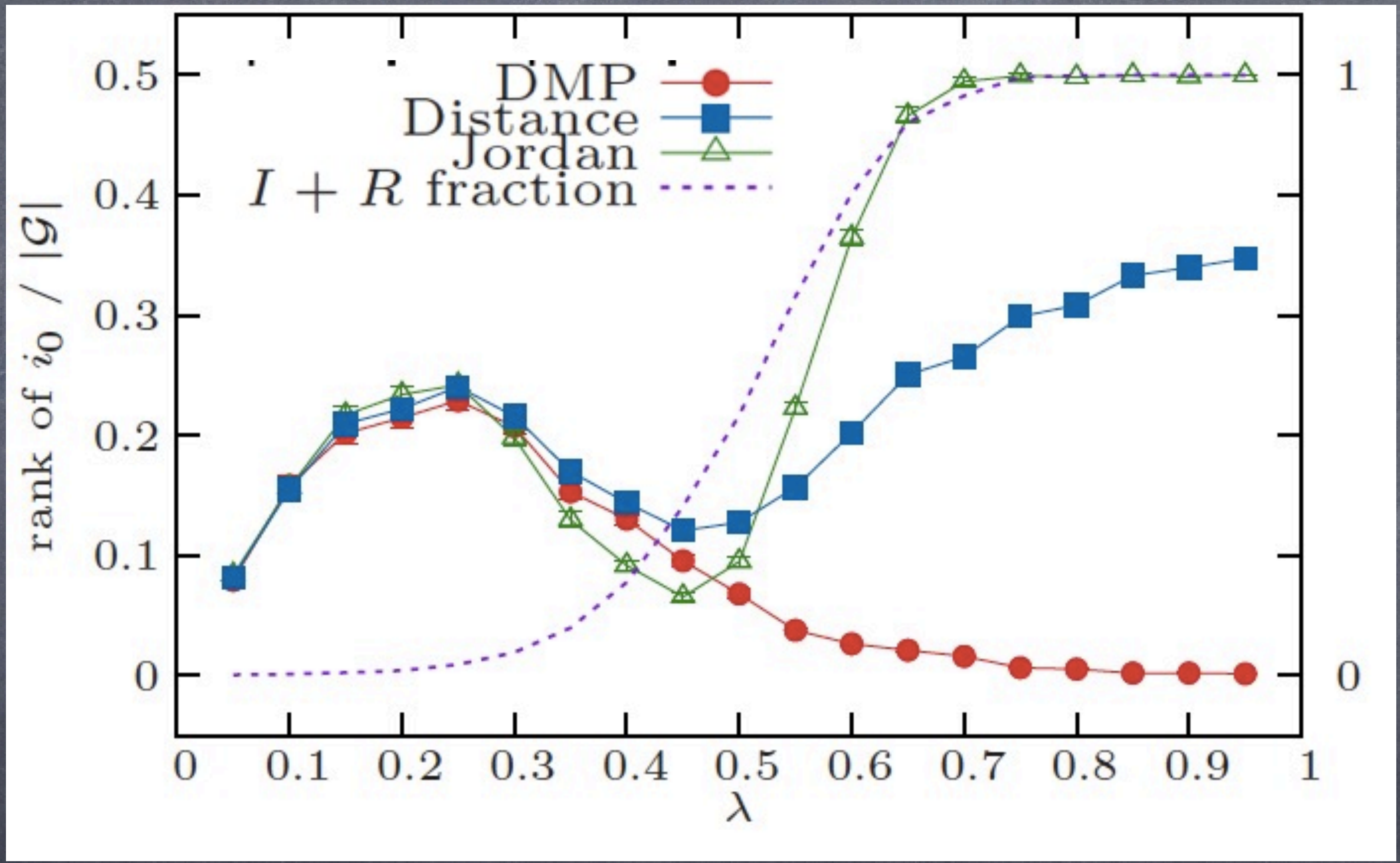
$\mu = 1, \lambda = 0.5$

$t_0 = 5$

1000 random instances

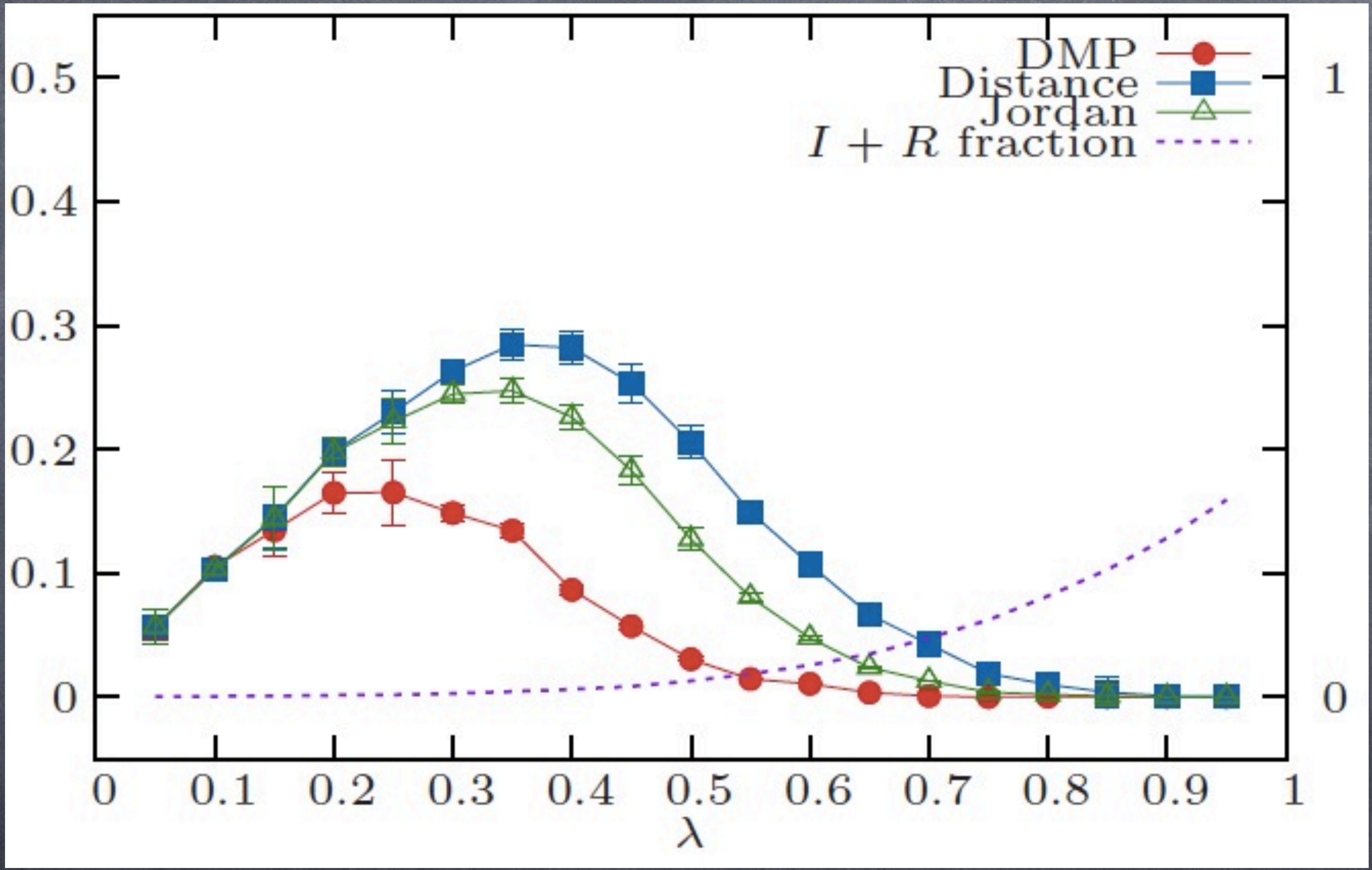


Results



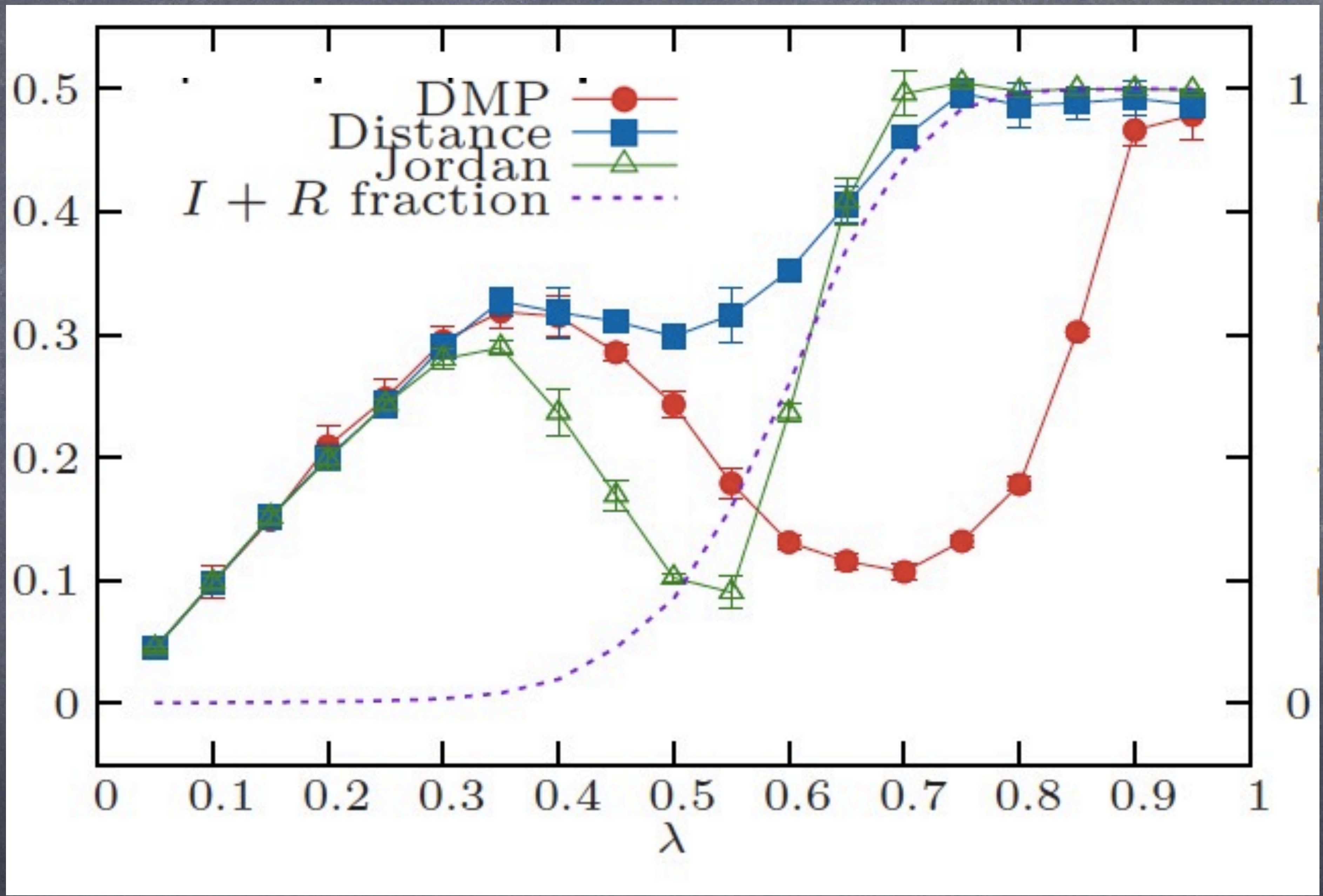
N=1000, c=4, $t_0 = 10, \mu = 0.5$

Results



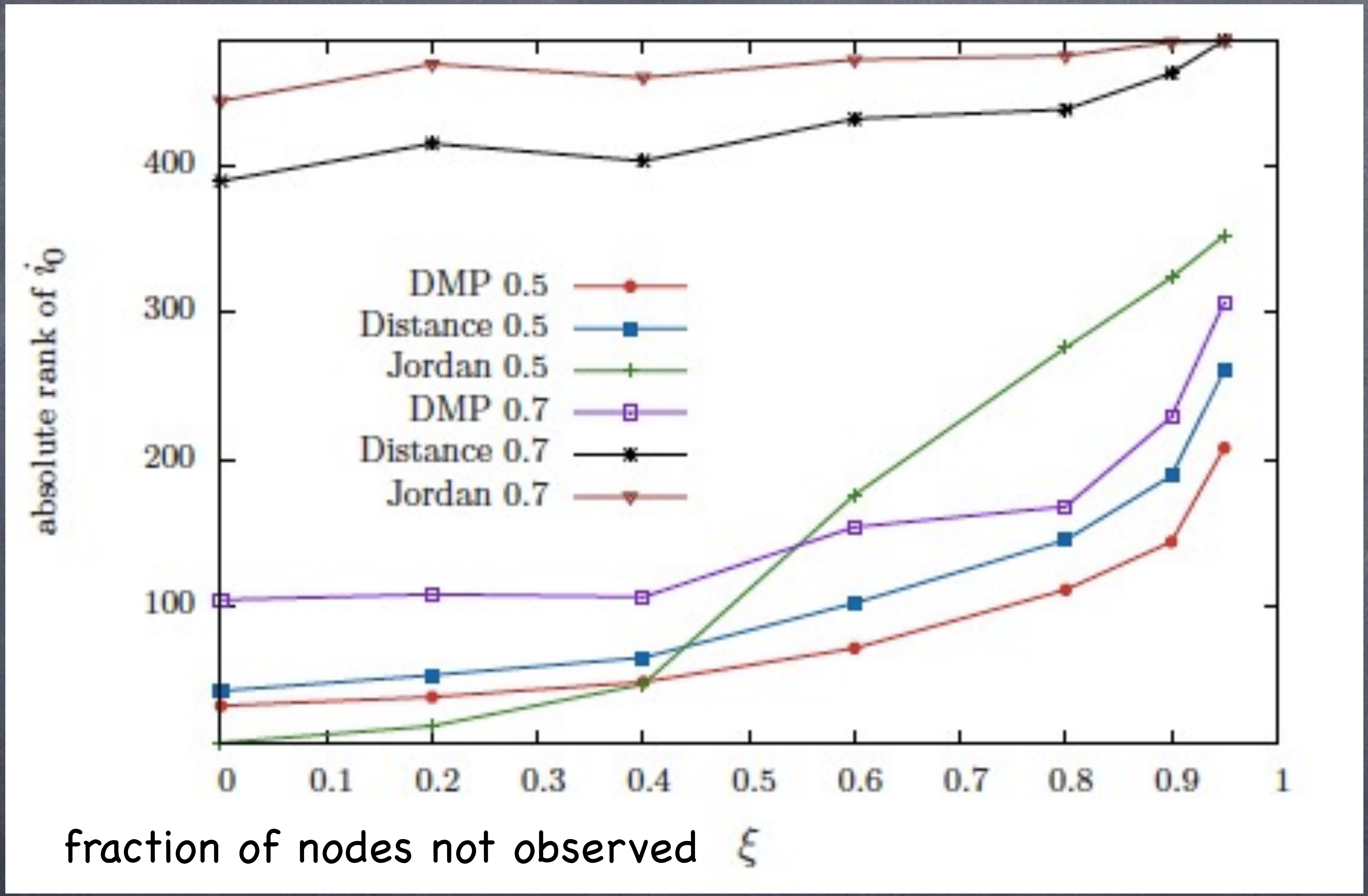
N=1000, c=4, $t_0 = 5, \mu = 1$

Results



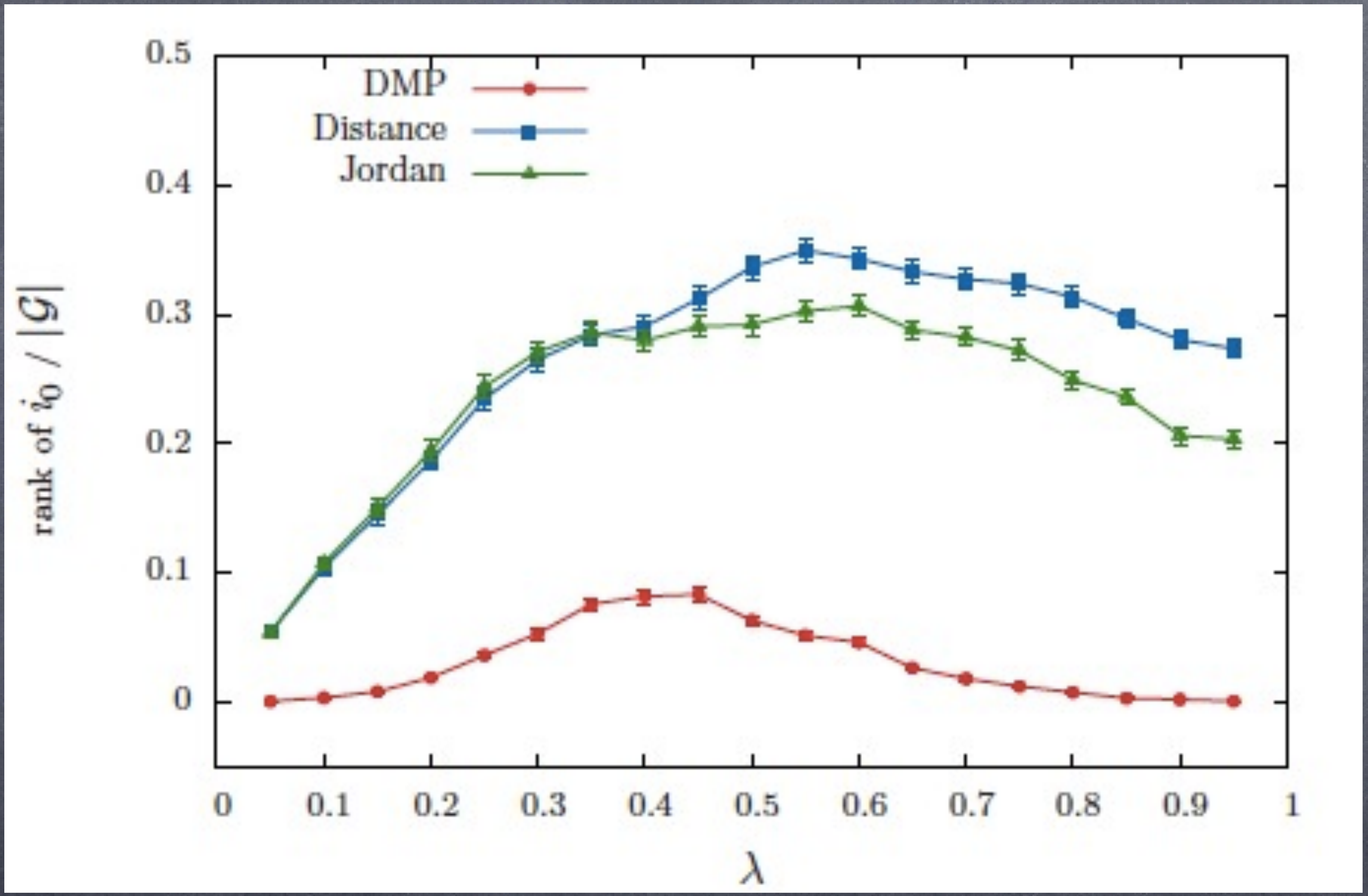
N=1000, c=4, $t_0 = 10, \mu = 1$

Results



N=1000, c=4, $t_0 = 10, \mu = 1$

Results



$t_0 = 10, \mu = 0.5$

U.S. East coast power grid, N=4941

Concluding remarks

- DMP improves in most situations over the various centralities.
- Our results show that inference of the origin is an relatively hard problem.
- Possible improvements: Incomplete knowledge of the graph. Better approximations for the likelihood than the mean field like (including pair-correlations did not improve results).
- Important general open question: For what other models is dynamical message passing tractable?

Concluding remarks

- DMP improves in most situations over the various centralities.
- Our results show that inference of the origin is an relatively hard problem.
- Possible improvements: Incomplete knowledge of the graph. Better approximations for the likelihood than the mean field like (including pair-correlations did not improve results).
- Important general open question: For what other models is dynamical message passing tractable?

Thank you!