



Recommendation Systems

part 3

School for advanced sciences of Luchon 2015

Debora Donato

debora@stumbleupon.com





- Friend Recommendation (a.k.a. Link prediction) and Network reconstructions
- Practical application and comparison of Local, Quasi-local, Global metrics (presented yesterday)
- Romantic partnerships and the dispersion of social ties



- Kristina Lerman
 - The Link-Prediction Problem for Social Networks (Liben-Nowell & Kleinberg)
- Yuan Shi
 - Link prediction in complex networks: a survey (L Lu and T Zhou)
 - Romantic Partnerships and the Dispersion of Social Ties (Backstrom & Kleinberg)



How You're Connected



You



Ana-Maria Popescu



Sudeep Das, Ph.D.



Kanishka Bhaduri

and 44 more connections in
common

[Get introduced](#)



Amanda Papp

Chi seguire · [Aggiorna](#) · [Visualizza tutto](#)



Dyaa Albakour @dyaaa

Seguito da [raffaele_perego](#) e...



Segui



Cool Scala @coolscala



Segui



ecir2016 @ecir2016

Seguito da [raffaele_perego](#) e...



Segui



Trova amici

Importa i tuoi contatti da Yahoo

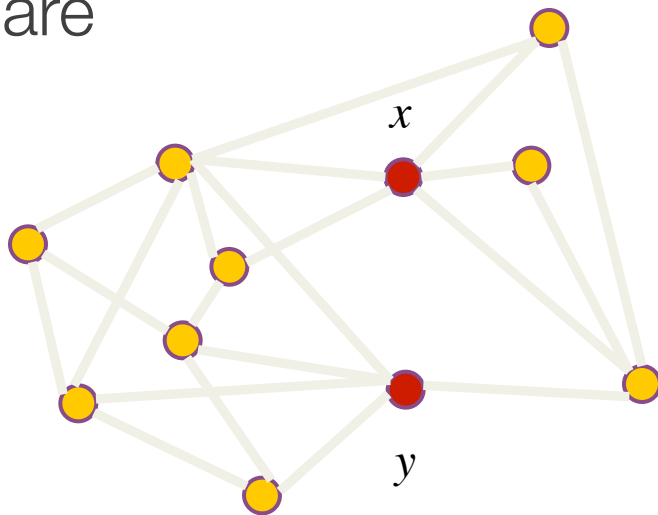
[Collega altre rubriche](#)

To what extent can the evolution of a social network be modeled using features *intrinsic to the network itself*?

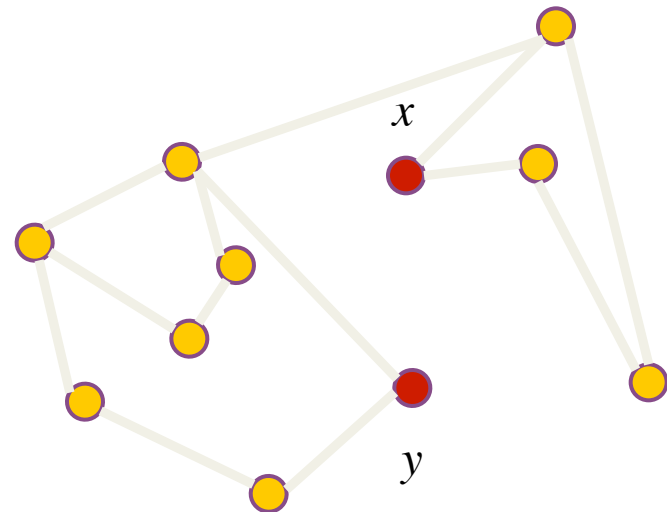
- Formalize the link prediction problem
 - Given a snapshot of a network, infer which new interactions between nodes are likely to occur in the future
- Propose link prediction heuristics based on measures for analyzing the “proximity” of nodes in a network.
- Evaluate link prediction heuristics on large coauthorship networks. Future coauthorships can be extracted from network topology.



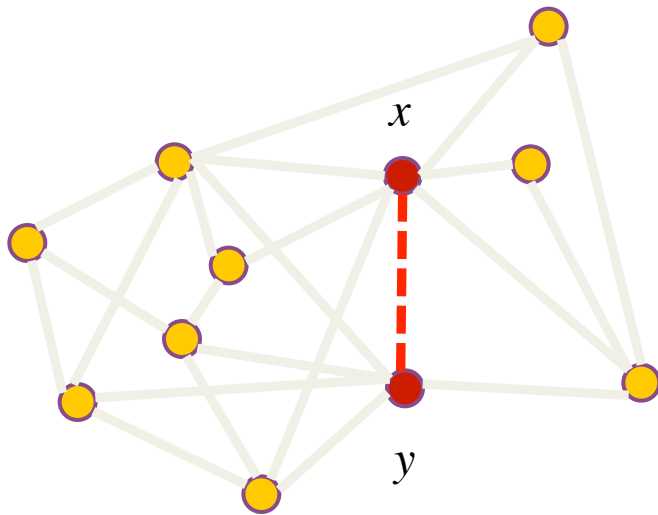
- In many networks, people who are “close” belong to the same social circles and will inevitably encounter one another and become linked themselves.
- Link prediction heuristics measure how “close” people are



Red nodes are close to each other



Red nodes are more distant

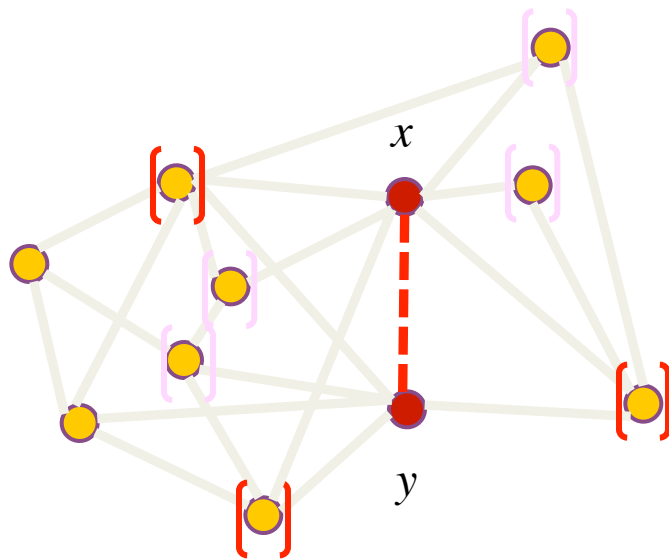


• Local

- Common neighbors (CN)
- Jaccard (JC)
- Adamic-Adar (AA)
- Preferential attachment (PA) ...

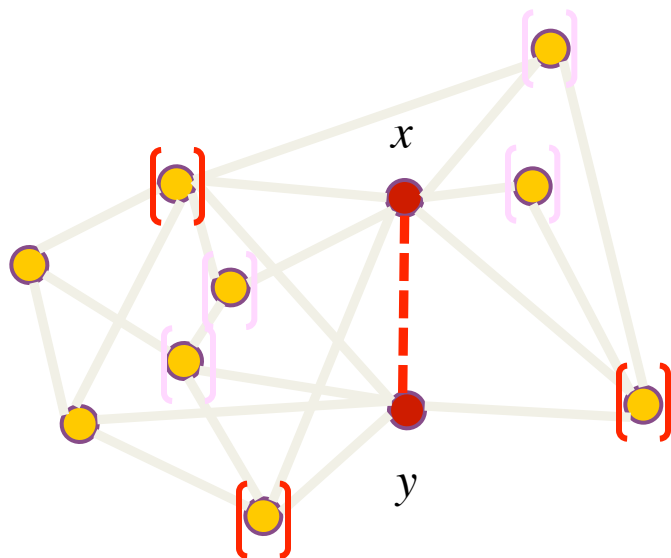
• Global

- Katz score
- Hitting time
- PageRank ...



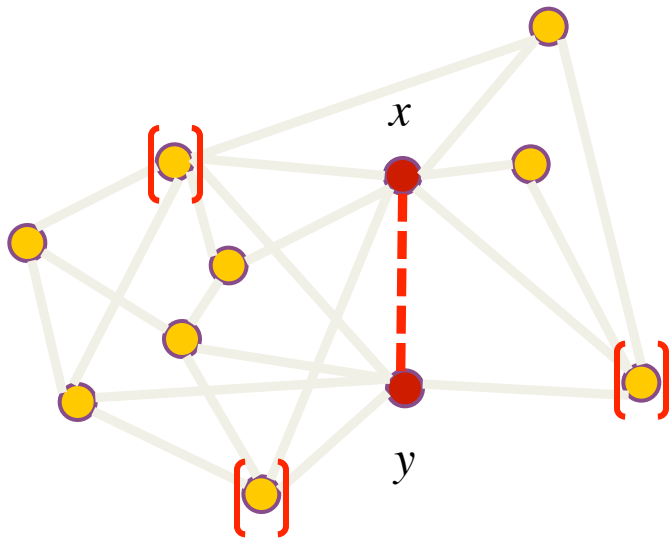
$$CN = 3$$

- Link prediction heuristics
 - Common neighbors (CN)
 - Neighborhood overlap
 - Jaccard (JC)
 - Adamic-Adar (AA)
 - Preferential attachment (PA)



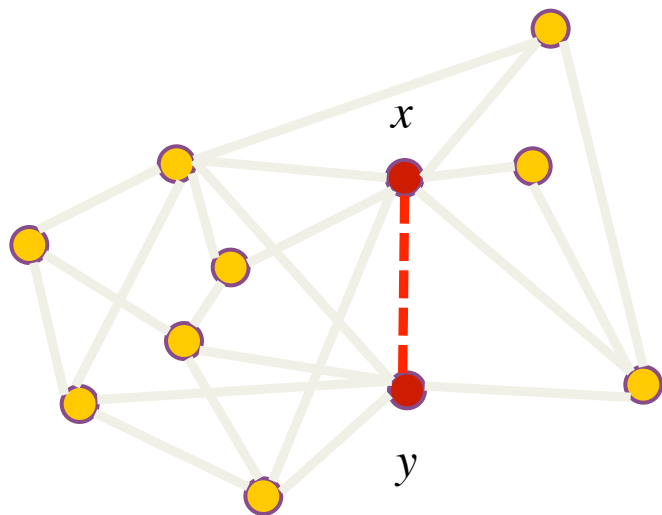
$$JC = \frac{CN}{d_x + d_y - CN} = 0.75$$

- Link prediction heuristics
 - Common neighbors (CN)
 - Jaccard (JC)
 - Fraction of common neighbors
 - Adamic-Adar (AA)
 - Preferential attachment (PA)



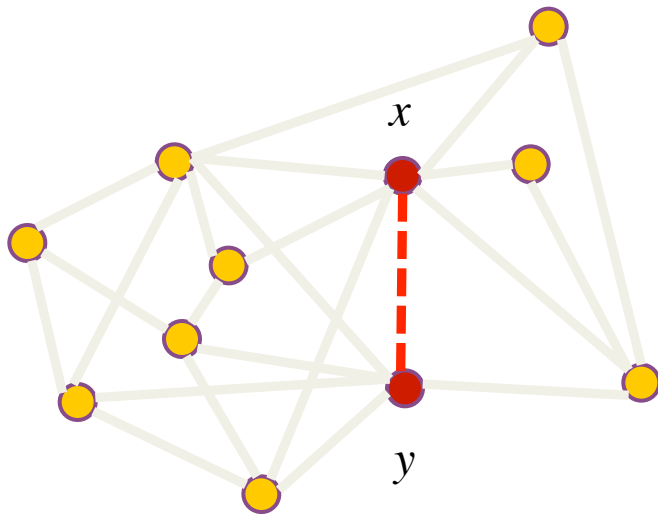
$$AA = \sum_{z \in CN} \frac{1}{\log d_z} = 4.6$$

- Link prediction heuristics
 - Common neighbors (CN)
 - Jaccard (JC)
 - Adamic-Adar (AA)
 - Nmbr common neighbors, with each neighbor z attenuated by log of its degree
 - Preferential attachment (PA)



$$PA = d_x d_y = 30$$

- Link prediction heuristics
 - Common neighbors (CN)
 - Jaccard (JC)
 - Adamic-Adar (AA)
 - Preferential attachment (PA)
 - Better connected nodes are more likely to form more links



• Link prediction heuristics

– Katz score

- Measures number of paths between two nodes, attenuated by their length

– Hitting time

- Expected time for a random walk from x to reach y

– ...



- Collaboration networks of physicists
 - Core nodes: authors who published at least 3 papers during the training period and at least 3 papers during test period
- Training data: graph $G(t_0, t_0')$ of collaborations during time period $[t_0, t_0']$ with V core nodes and E_{old} edges
- Test data: graph $G(t_1, t_1')$ of collaborations during a later time period $[t_1, t_1']$ with V core nodes and E_{new} edges

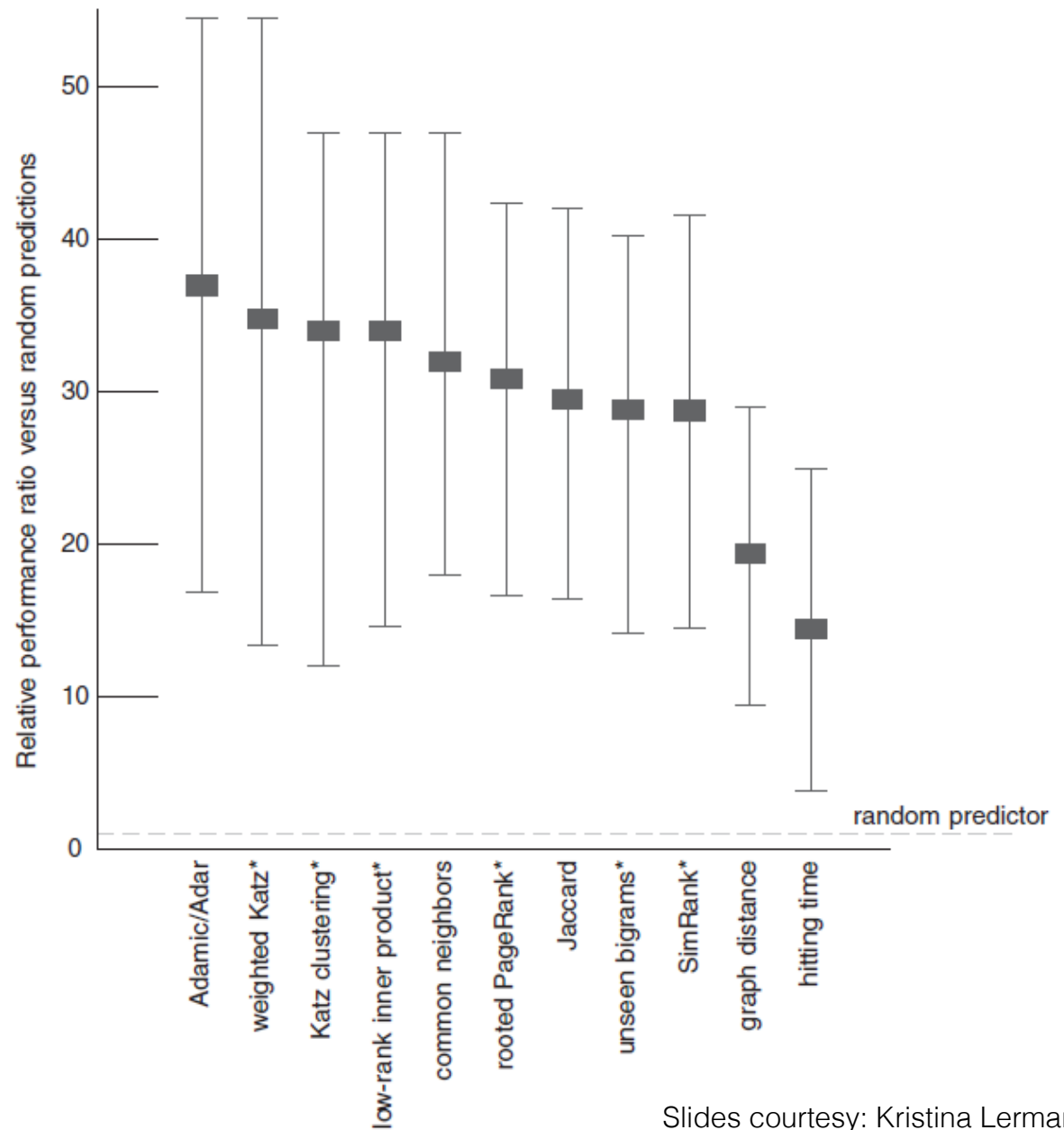
	Training Period			Core		
	Authors	Articles	Collaborations ^a	Authors	E_{old}	E_{new}
astro-ph	5,343	5,816	41,852	1,561	6,178	5,751
cond-mat	5,469	6,700	19,881	1,253	1,899	1,150
gr-qc	2,122	3,287	5,724	486	519	400
hep-ph	5,414	10,254	47,806	1,790	6,654	3,294
hep-th	5,241	9,498	15,842	1,438	2,311	1,576



- Link prediction algorithm
 - Score node pairs using a heuristic p
 - New links more likely among high scoring pairs
- Each link prediction heuristic p outputs a ranked list L of new collaborations: pairs in $V \times V - E_{old}$.
- Focus evaluation on new links E_{new}^* between core nodes
- Performance metric: How many of the top n pairs in ranked list L are the actual new nodes in E_{new}^* ?

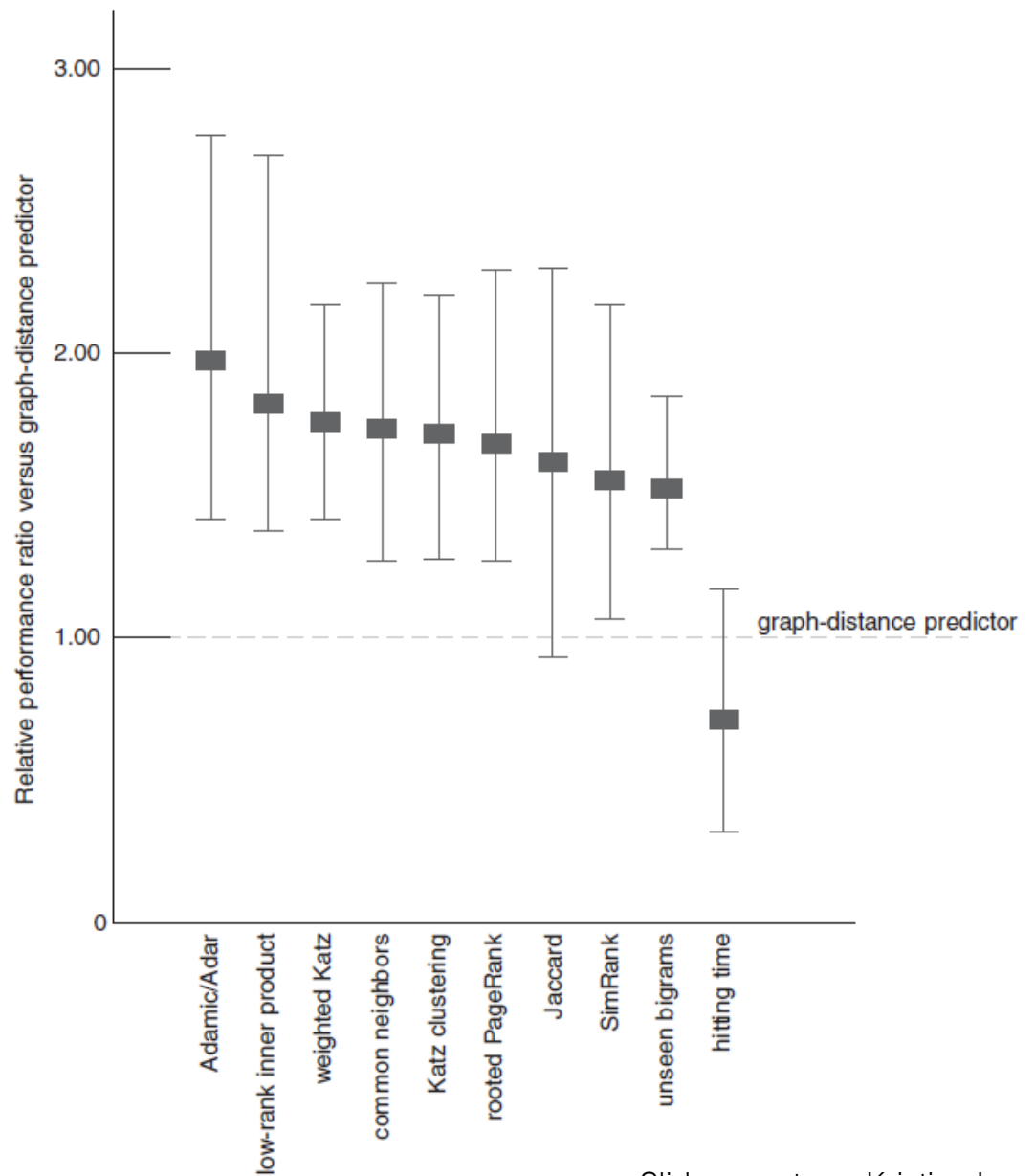


Heuristics vs random predictor





Heuristics vs graph distance predictor

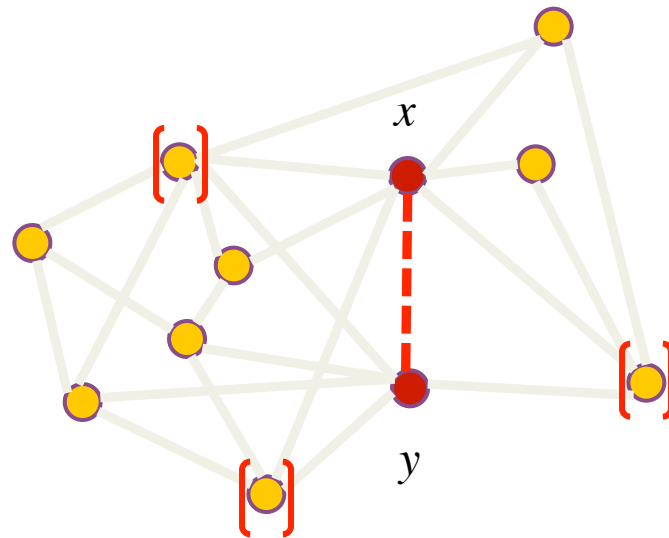




- Graph-based link prediction heuristics outperform random guess by a factor of ~ 40
 - Best performing AA and Katz
- Graph-based link prediction heuristics outperform graph-distance by a factor of ~ 2
 - Best performing AA and PA
- However, they still predict only 16% of new collaborations at best, leaving much room for improvement.



- Extensive reviews of all the indexes and experimentation
 - 10 local indexes
 - 7 global indexes



$$RA = \sum_{z \in CN} \frac{1}{d_z} = 0.67$$



Metric: AUC. Each number averaged by 10 implementations.

Real-world networks

PPI: protein-protein interaction

NS: co-authorship

Grid: electrical power-grid

PB: US political blogs

INT: router-level Internet

USAir: US air transportation

CN and AA have second best performance

Indices	PPI	NS	Grid	PB	INT	USAir
CN	0.889	0.933	0.590	0.925	0.559	0.937
Salton	0.869	0.911	0.585	0.874	0.552	0.898
Jaccard	0.888	0.933	0.590	0.882	0.559	0.901
Sørensen	0.888	0.933	0.590	0.881	0.559	0.902
HPI	0.868	0.911	0.585	0.852	0.552	0.857
HDI	0.888	0.933	0.590	0.877	0.559	0.895
LHN1	0.866	0.911	0.585	0.772	0.552	0.758
PA	0.828	0.623	0.446	0.907	0.464	0.886
AA	0.888	0.932	0.590	0.922	0.559	0.925
RA	0.890	0.933	0.590	0.931	0.559	0.955

RA performs the best



- Local Path Index (LP):

$$S^{LP} = A^2 + \epsilon A^3$$

- Local Random Walk (LRW): at time step t ,

$$s_{xy}^{LRW}(t) = q_x \pi_{xy}(t) + q_y \pi_{yx}(t)$$

$$\vec{\pi}_x(0) = \vec{e}_x \quad \vec{\pi}_x(t+1) = P^T \vec{\pi}_x(t) \text{ for } t \geq 0$$

- Superposed Random Walk (SRW): at time step t ,

$$s_{xy}^{SRW}(t) = \sum_{\tau=1}^t s_{xy}^{LRW}(\tau) = \sum_{\tau=1}^t [q_x \pi_{xy}(\tau) + q_y \pi_{yx}(\tau)]$$



- Katz Index:

$$s_{xy}^{Katz} = \sum_{l=1}^{\infty} \beta^l \cdot |paths_{xy}^{<l>}| = \beta A_{xy} + \beta^2 (A^2)_{xy} + \beta^3 (A^3)_{xy} + \dots$$

- Random Walk with Restart (direct application of PageRank algorithm)



Metric: AUC. Each number averaged by 10 implementations.

Real-world networks

PPI: protein-protein interaction

NS: co-authorship

Grid: electrical power-grid

PB: US political blogs

INT: router-level Internet

USAir: US air transportation

AUC	PPI	NS	Grid	PB	INT	USAir
LP	0.970	0.988	0.697	0.941	0.943	0.960
LP*	0.970	0.988	0.697	0.939	0.941	0.959
Katz	0.972	0.988	0.952	0.936	0.975	0.956
LHN2	0.968	0.986	0.947	0.769	0.959	0.778
Precision	PPI	NS	Grid	PB	INT	USAir
LP	0.734	0.292	0.132	0.519	0.557	0.627
LP*	0.734	0.292	0.132	0.469	0.121	0.627
Katz	0.719	0.290	0.063	0.456	0.368	0.623
LHN2	0	0.060	0.005	0	0	0.005

Katz performs the best →

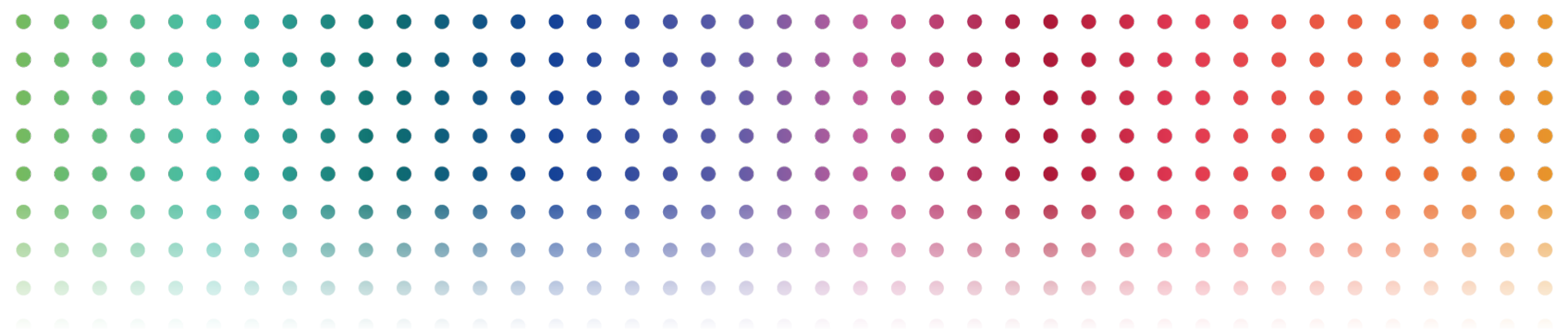


Metric: AUC. Each number averaged by 10 implementations.

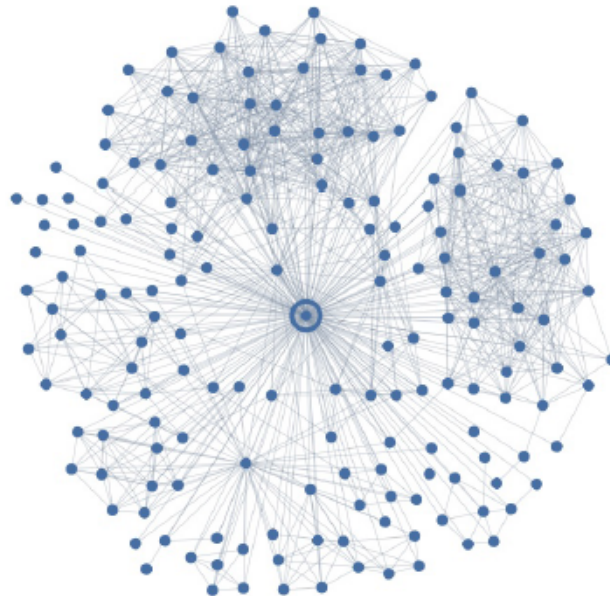
AUC	CN	RA	LP	ACT	RWR	HSM	LRW	SRW
USAir	0.954	0.972	0.952	0.901	0.977	0.904	0.972(2)	0.978(3)
NetScience	0.978	0.983	0.986	0.934	0.993	0.930	0.989(4)	0.992(3)
Power	0.626	0.626	0.697	0.895	0.760	0.503	0.953(16)	0.963(16)
Yeast	0.915	0.916	0.970	0.900	0.978	0.672	0.974(7)	0.980(8)
C.elegans	0.849	0.871	0.867	0.747	0.889	0.808	0.899(3)	0.906(3)
Precision	CN	RA	LP	ACT	RWR	HSM	LRW	SRW
USAir	0.59	0.64	0.61	0.49	0.65	0.28	0.64(3)	0.67(3)
NetScience	0.26	0.54	0.30	0.19	0.55	0.25	0.54(2)	0.54(2)
Power	0.11	0.08	0.13	0.08	0.09	0.00	0.08(2)	0.11(3)
Yeast	0.67	0.49	0.68	0.57	0.52	0.84	0.86(3)	0.73(9)
C.elegans	0.12	0.13	0.14	0.07	0.13	0.08	0.14(3)	0.14(3)



- Global indices
 - Pros: more accurate than local indices
 - Cons: 1) time-consuming; 2) global topological information may not be available
- Local Index
 - LRW and SRW best performing

A decorative header consisting of a grid of colored dots in shades of green, blue, purple, red, and orange, arranged in a pattern that tapers off to the right.

Romantic partnerships and the dispersion of social ties



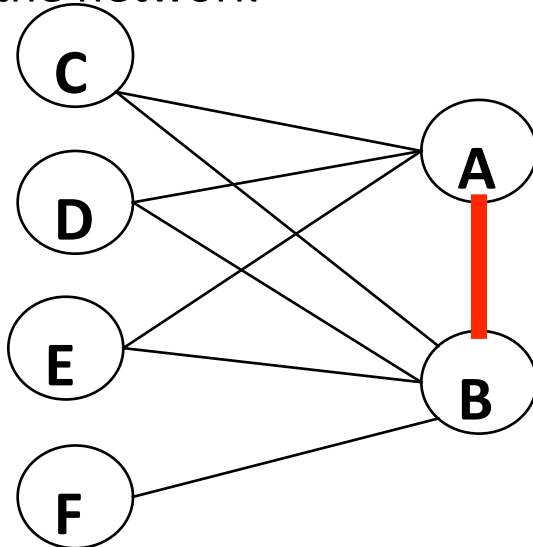
- Questions
 - Who are the most important individuals in a person's social neighborhood?
 - What are the defining structural signatures of a person's social neighborhood?
- Contributions
 - Dispersion: a new measure for estimating tie strength
 - Characterize romantic relationships in terms of network structure
 - Empirical study of this characteristic across Facebook population



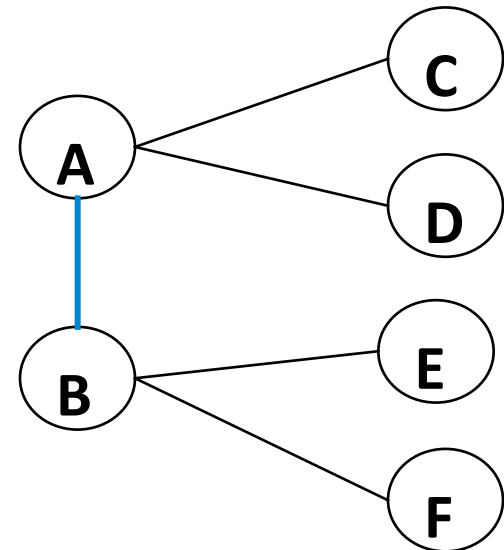
Who are the most important people in one's social neighborhood?

- Following Granovetter, researchers use number of mutual friends (embeddedness) to identify strong ties
 - Close friends, who share much time together
 - Emotionally intense interactions

A-B tie is highly embedded in the network



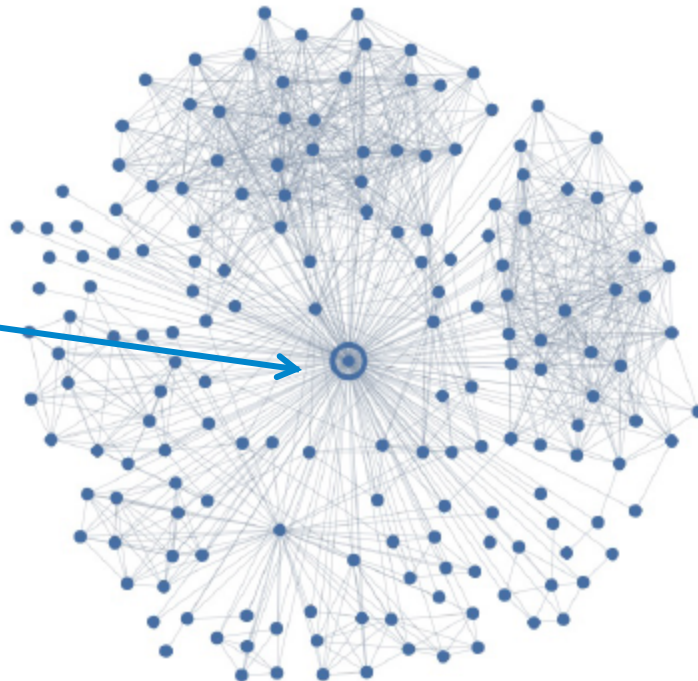
A-B tie is not embedded in the network





- Embeddedness is not able to identify “significant others” (romantic relationships, e.g., spouse, partner, boy/girlfriend)
- Ego network – social neighborhood of an individual, showing all his/her friends and links between them

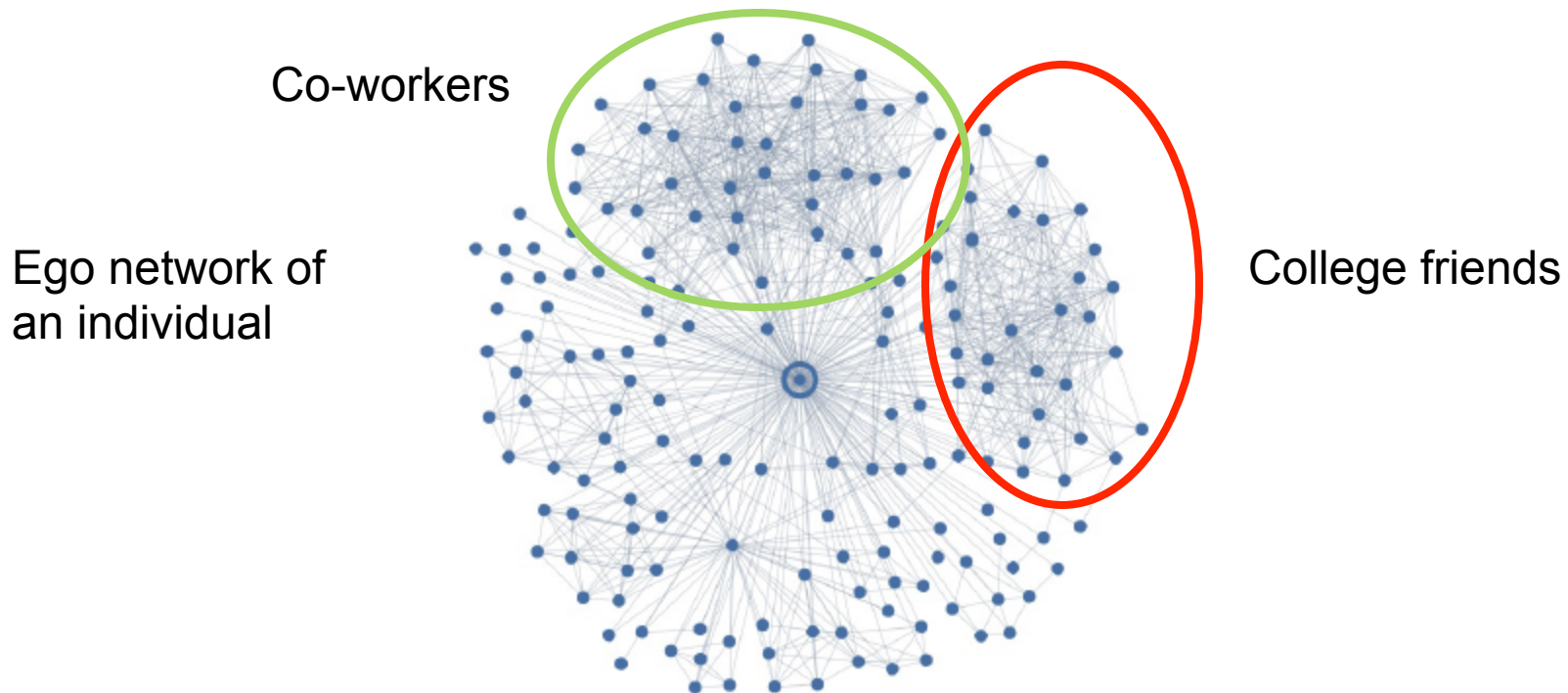
Ego network of
an individual



**Who is the
“significant other”?**



- People have large clusters of friends corresponding to well-defined foci of interaction in their lives
 - These links have high embeddedness but are not very strong ties
- In contrast, romantic partners may have lower embeddedness, but they often involve mutual friends from different foci

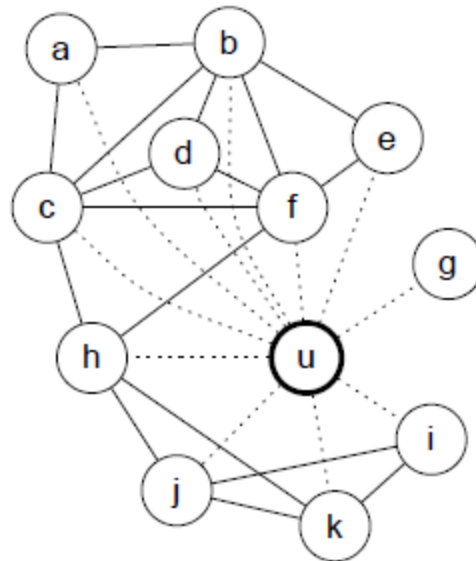




Embeddedness:
u and v have many mutual neighbors.

Links u-b, u-c, and u-f have embeddedness 5

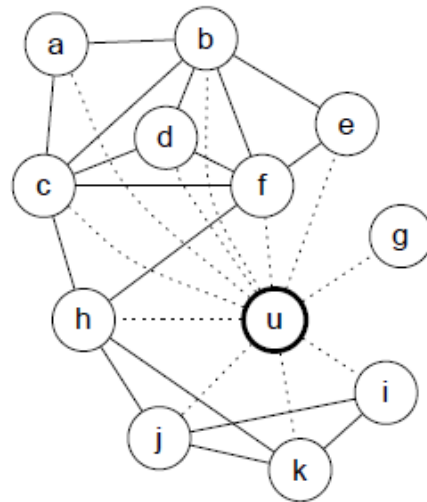
Link u-h has embeddedness 4



Dispersion:

mutual neighbors of u and v are not well-connected to one another, and hence u and v are the only intermediaries between these different parts of the network.

Link u-h has high dispersion: u and h are the only intermediaries between c and f



- Let C_{uv} be the set of common neighbors of u and v

$$disp(u, v) = \sum_{s, t \in C_{uv}} d_v(s, t)$$

- $d(s, t)$ is the distance between s and t .
 - For simplicity, take $d(s, t) = 1$ when s, t are not directly linked and have no common neighbors in the egonet, other than u and v

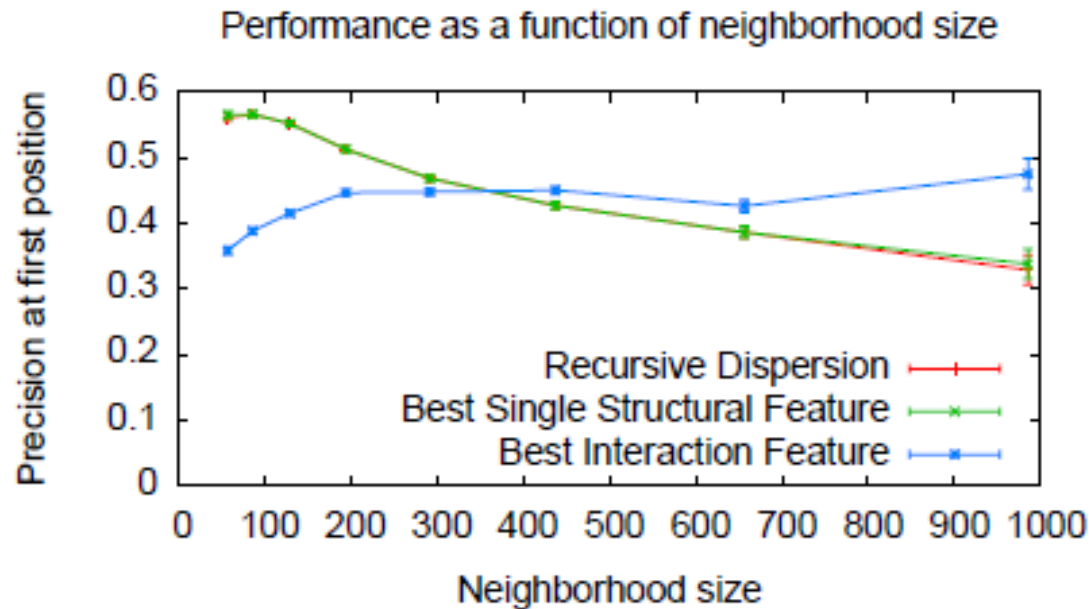


- Egonetworks of 1.3 million Facebook users, selected uniformly at random from among all users of age at least 20, with between 50 and 2000 friends, who list a spouse or relationship partner in their profile
- Rank all friends by importance. Attempt to identify romantic partners
- Measure: Precision of the first position, $\text{Pr}@1$



- How well does dispersion predict the “significant other”? – precision of the top-ranked person in the individual’s egonet
 - Beats others measures of interaction between users
 - viewing of profiles, sending of messages, and co-presence at events (photos)

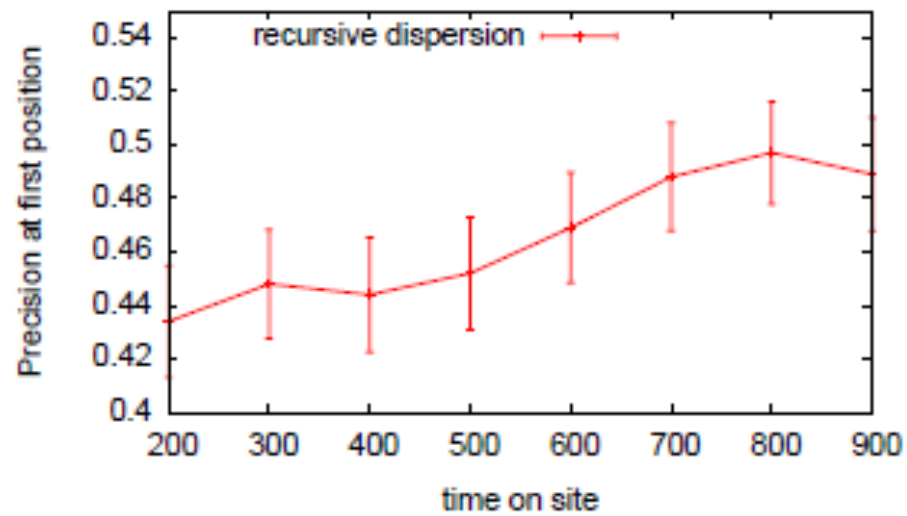
type	embed	rec.disp.	photo	prof.view.
all	0.247	0.506	0.415	0.301
married	0.321	0.607	0.449	0.210
married (fem)	0.296	0.551	0.391	0.202
married (male)	0.347	0.667	0.511	0.220
engaged	0.179	0.446	0.442	0.391
engaged (fem)	0.171	0.399	0.386	0.401
engaged (male)	0.185	0.490	0.495	0.381
relationship	0.132	0.344	0.347	0.441
relationship (fem)	0.139	0.316	0.290	0.467
relationship (male)	0.125	0.369	0.399	0.418



- Performance is best when the neighborhood size is around 100 nodes (56%), & drops moderately (to 33%) as the egonet size increases by an order of magnitude to 1000
- Interaction features are better for larger neighborhoods, due to users with larger neighborhoods being more active



Performance as a function of user's time on site



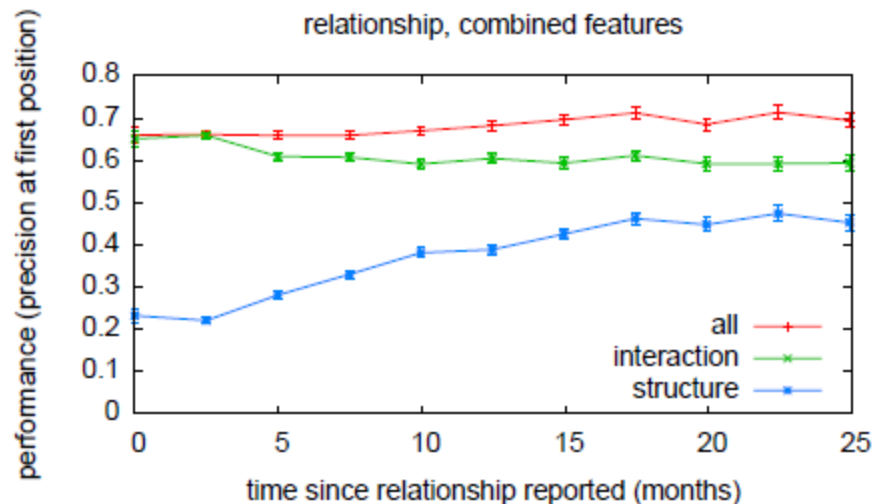
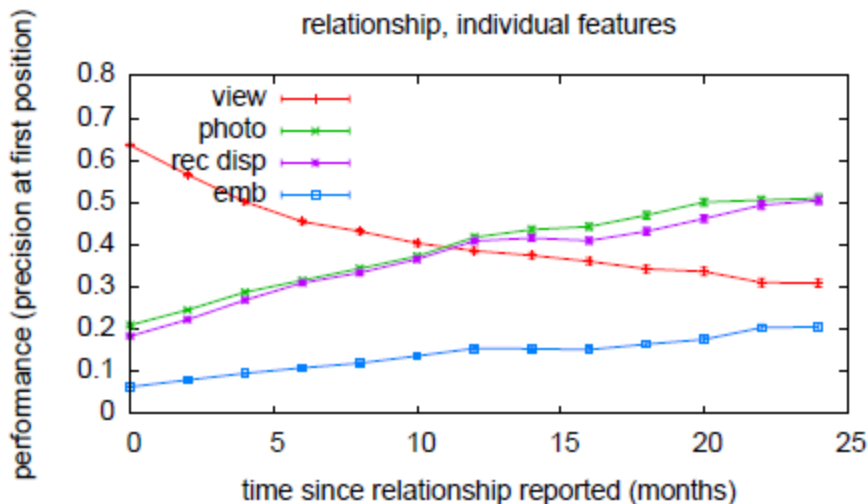
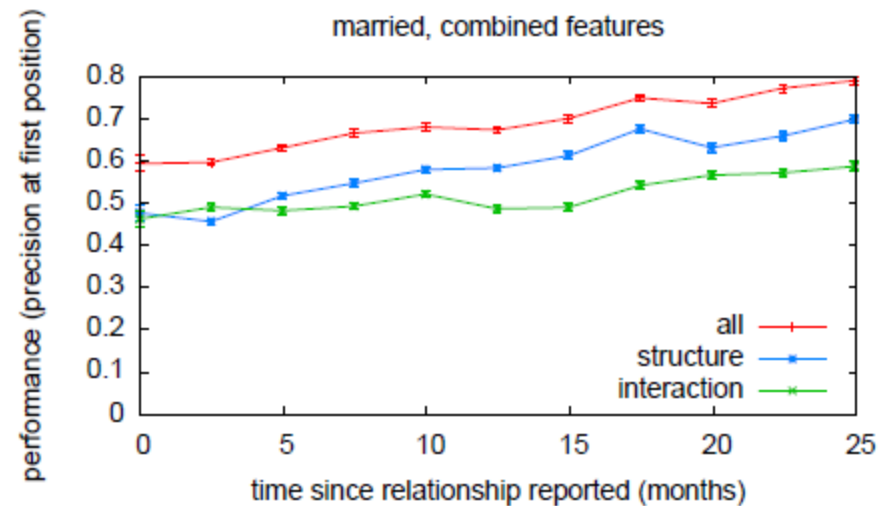
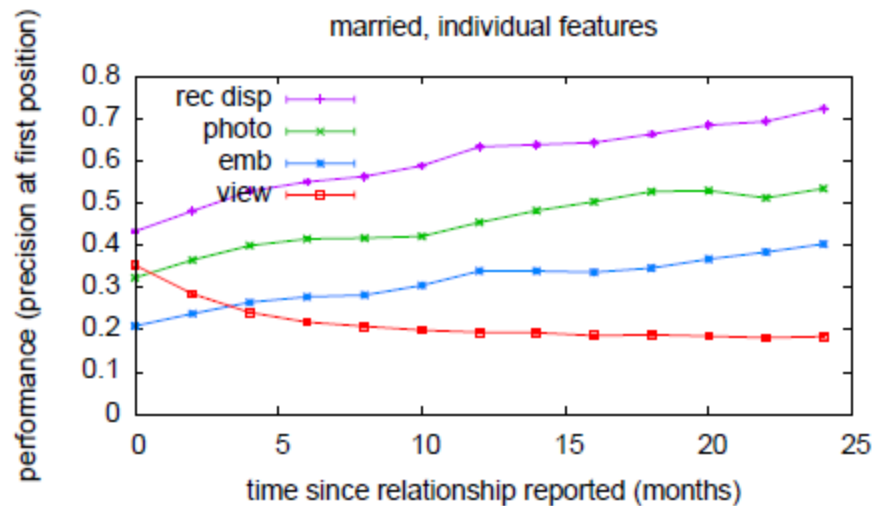


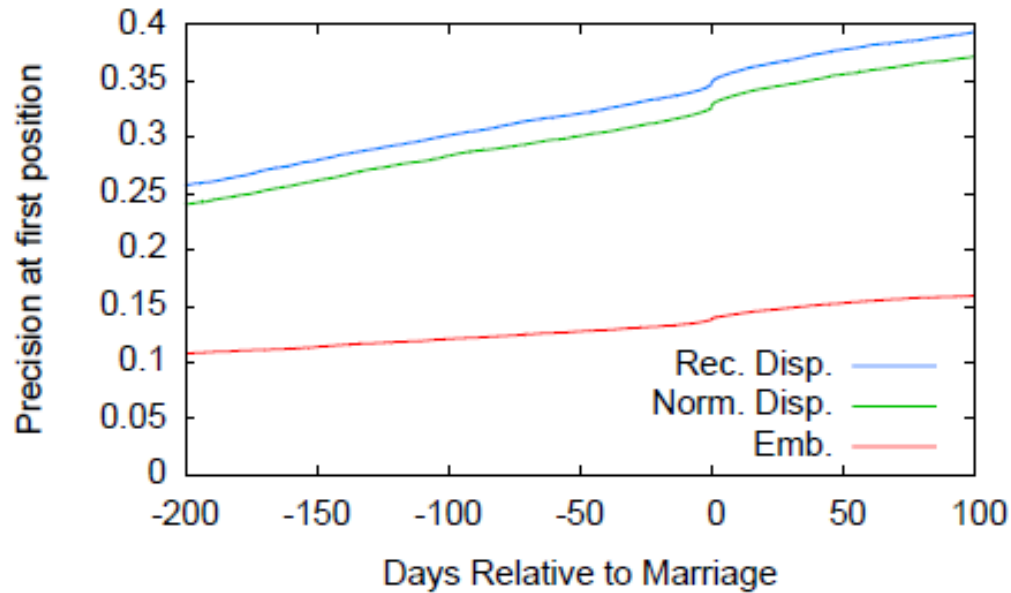
Task	baseline	demo.	network	both
Single vs. Any Rel.	59.8%	67.9%	61.6%	68.3%
Single vs. Married	56.6%	78.0%	66.1%	79.0%

- Predict relationship status of users
 - Ground truth: 60% of users are in a relationship
- Demographic features (age, gender, country, and time on site) work better than network-based features (dispersion)
- Best performance combining demographic and network features

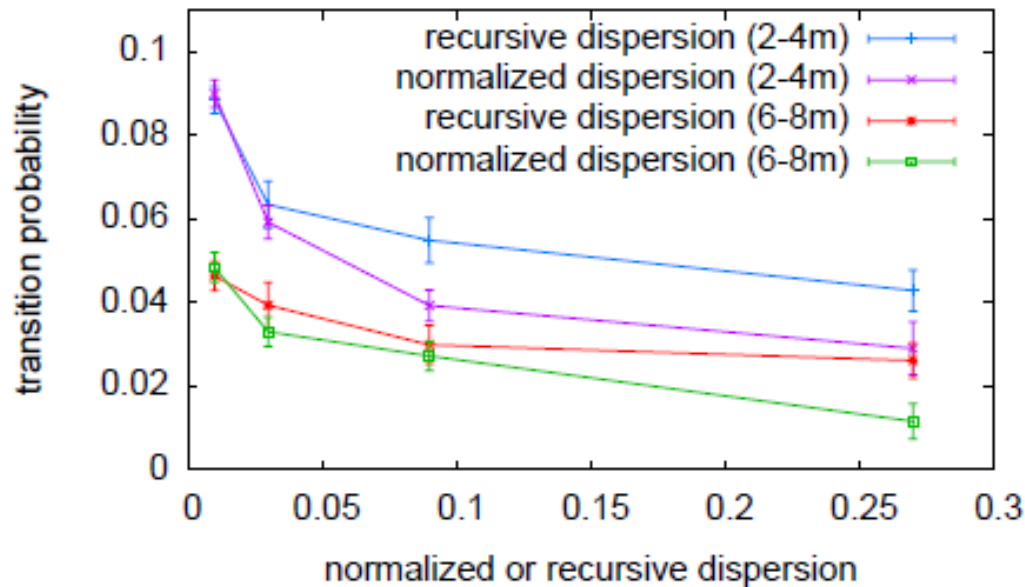


How does performance vary based on age of the relationship?





- Performance of dispersion measures increases as people approach time of their marriage



- Transition probability from the status ‘in a relationship’ to the status ‘single’ over a 60-day period. The transition probabilities decrease monotonically, and by significant factors, for users with high normalized or recursive dispersion to their respective partners.



- Graph structure contains information predictive of individual relationships
 - New collaborations
 - Romantic partnerships
- In many cases, graph-based algorithms outperform feature-based machine learning algorithms
- These suggest complex interactions between personal relationships and global network structure