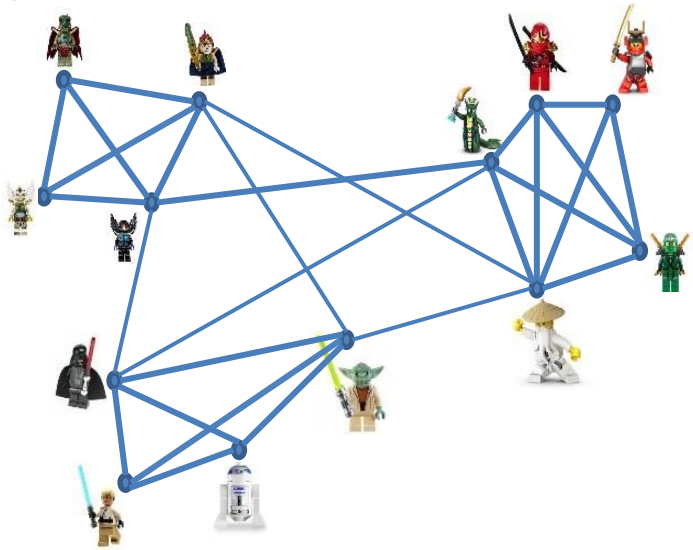
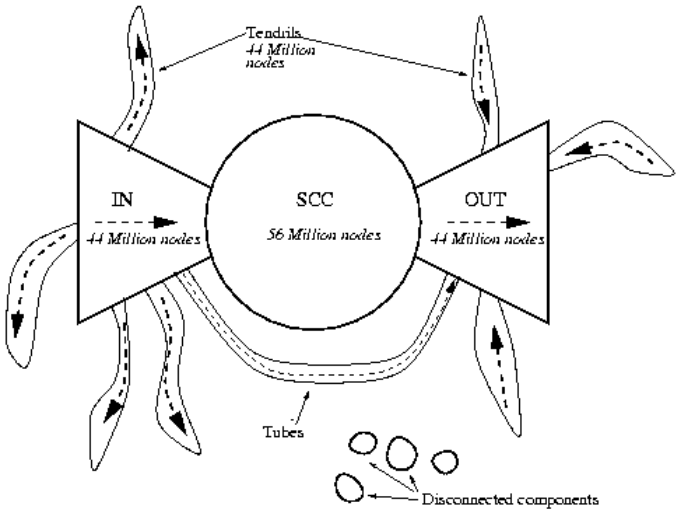
Scalable Mining of Massive Networks: Distance-based Centrality, Similarity, and Influence

Edith Cohen

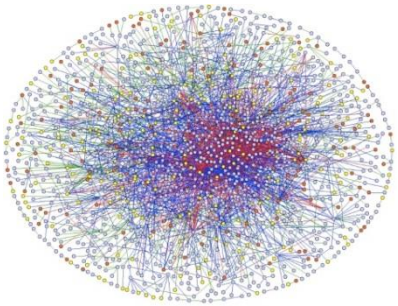
Tel Aviv University

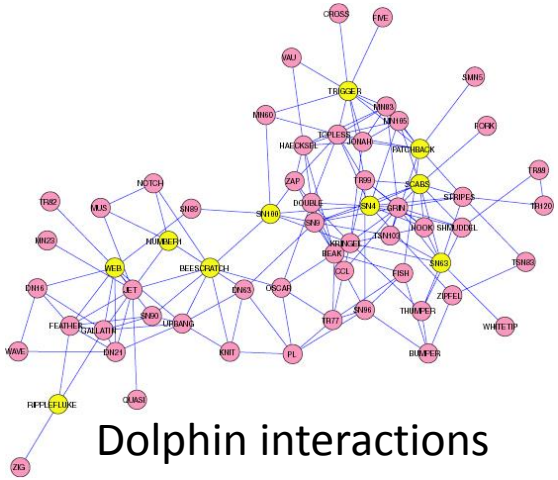


Graph Datasets:
Represent relations between “things”



Bowtie structure of the Web Broder et. al. 2001





Dolphin interactions

Graph Datasets

- Hyperlinks (the Web)
- Social graphs (Facebook, Twitter, LinkedIn,…): friend, follow, like
- Email logs, phone call logs , messages
- Commerce transactions (Amazon,eBay)
- Road networks
- Communication networks
- Protein interactions
- …

Analytics on Graphs

Centralities/Influence

- The power/importance/coverage of a node or a set of nodes
- Applications: ranking, viral marketing,…

Similarities/Communities

- How tightly related are 2 or more nodes
- Applications: Friend/Product Recommendations, attribute completion, Advertising, Prediction

Challenges

- **Modeling:**
  - Models must capture “reality”
  - Be flexible to allow fitting to problem (tunable parameters)
- **Scalability:**
  - 10^{9+} interactions (edges), 10^{8+} entities (nodes)

Centrality, Similarity, Influence

Structural measures (based on the set of interactions)

Local measures: only depend on the set of neighbors (paths of length 1)

- Degree centrality ,
- Influence as cardinality of union of neighbors
- Similarity by relation of neighbor sets (Jaccard, Adamic/Adar)

Advantages:

- Gets the “first order bit”
- Scalable

Disadvantages:

- Limited recall
- Spammable

Centrality, Similarity, Influence

Structural measures (based on the set of interactions)

Global Measures: depend on the paths ensemble

Higher recall but much less scalable

- Random walk based (Geometric/Heat Kernel): Centrality, similarity, influence thru (personalized) page rank
- Katz (Sum of paths, discounted by length)
- Resistance: Similarity by effective resistance
- Distance/reachability based Measures: centrality, similarity, influence

Distance/Reachability based measures of Centrality, Influence, Similarity

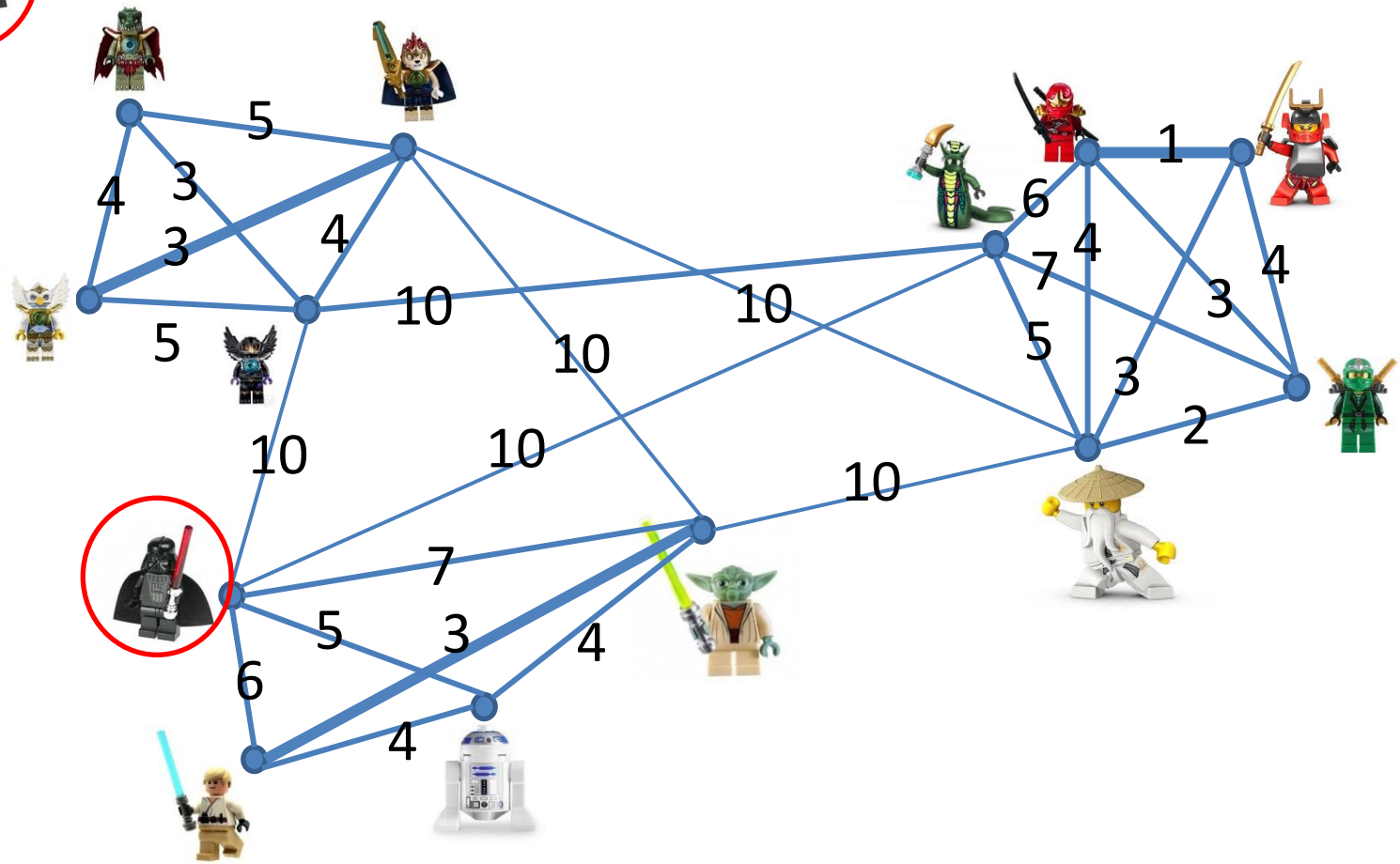
- ***Closeness centrality:*** Ability of a node to "reach" other nodes [Bavelas 1950]+++:
- ***Influence:*** Ability of a set of nodes to reach others [GLM2001,KKT2003,] [Gomez-Rodriguez+ ICML 2011], [Du+ NIPS 2013]. [C'DPW 2014] +++:
- ***Closeness similarity:*** Relates two nodes based on the similarity of their "reach" [C'DFGGW2013], (local: [AA 2005,LK2007]):

This talk: Unified treatment of distance/reachability-based measures

- Models:
  - Basic definition of “distance-based”
  - Different ways of using distance/reachability
  - Argue that models capture intuitive properties
- Scalability: Sketching and estimation

Distance vectors

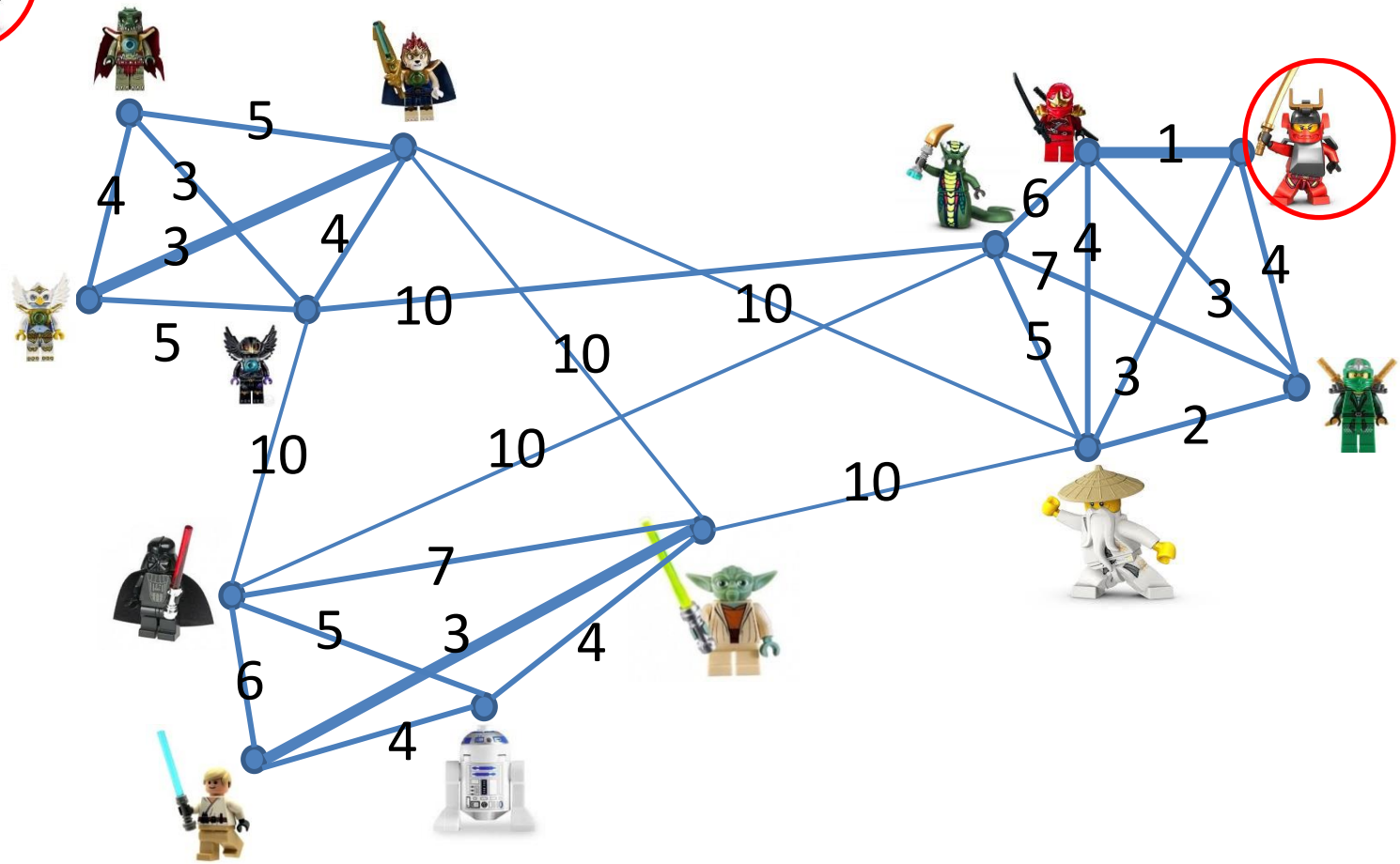
SP distances:



Distance vectors

SP distances:

17 17 16 13 7 17 16 13 16 3 1 0 4



Distance vectors

SP distances d_{ij}:

0 5 6 7 10 10 13 14 15 15 16 17 17

17 17 16 13 7 17 16 13 16 3 1 0 4

Distance-based measures

SP distances d_{ij}:

0	5	6	7	10	10	13	14	15	15	16	17	17

17	17	16	13	7	17	16	13	16	3	1	0	4

Relate entities based on their vectors.

- Closeness centrality $C($ $)$
- Closeness Similarity Sim(,)
- Influence Inf(, , ,)

Distance-based Models

SP distances d_{ij}:

	0	5	6	7	10	10	13	14	15	15	16	17	17
	17	17	16	13	7	17	16	13	16	3	1	0	4

- Inbound/outbound/undirected distance
- Distance/Reachability/Dijkstra (NN) rank
- Topic filter $\beta(v) \geq 0$: weighted relevance of entity
- Distances vs. decayed distances $\alpha(d) \geq 0$
- Robustness by randomizing lengths/presence of edges (REL), or using multiple “example” instances

Distance-based Models

SP distances d_{ij}:

	0	5	6	7	10	10	13	14	15	15	16	17	17
	17	17	16	13	7	17	16	13	16	3	1	0	4

- Inbound/outbound/undirected distance
- Distance/Reachability/Dijkstra (NN) rank
- Topic filter $\beta(v) \geq 0$: weighted relevance of entity
- Distances vs. decayed distances $\alpha(d) \geq 0$
- Robustness by randomizing lengths/presence of edges (REL), or using multiple “example” instances

Topic Filter $\boldsymbol{\beta(i)}$

Weigh entities (nodes or entries of vectors) according to

- Topic, interests, education level, age, community, geography, language, product type,…
- Applications: focus measure on topic. Useful for recommendations, attribute completion, targeted ads

➢ Similar to the role of "start state" in PPR

"Evil"

$\boldsymbol{\beta}$:	1	0	0	0.8	0.9	0	0.1

Distance-based Models

SP distances d_{ij}:

	0	5	6	7	10	10	13	14	15	15	16	17	17
	17	17	16	13	7	17	16	13	16	3	1	0	4

- Inbound/outbound/undirected distance
- Distance/Reachability/Dijkstra (NN) rank
- Topic filter $\beta(v) \geq 0$: weighted relevance of entity
- Distances vs. decayed distances $\alpha(d) \geq 0$
- Robustness by randomizing lengths/presence of edges (REL), or using multiple “example” instances

Farness-penalty vs. Closeness-reward

- To emphasize penalty for weak coverage of far nodes, we use distances as is: The norm of the distances vector of a node is dominated by "far" nodes. As in classic closeness centrality, facility location, k-medians/means
- To emphasize reward for better coverage of close nodes, we apply a decay ("kernel") function $\alpha(d)$ to the distance: The norm is dominated by closer nodes. As in influence models and local measures.

- Exact computation: both requires full distance vectors.
- Scalable approximation: different techniques

Quantifying Closeness Reward: Distance Decay (Kernel) Functions

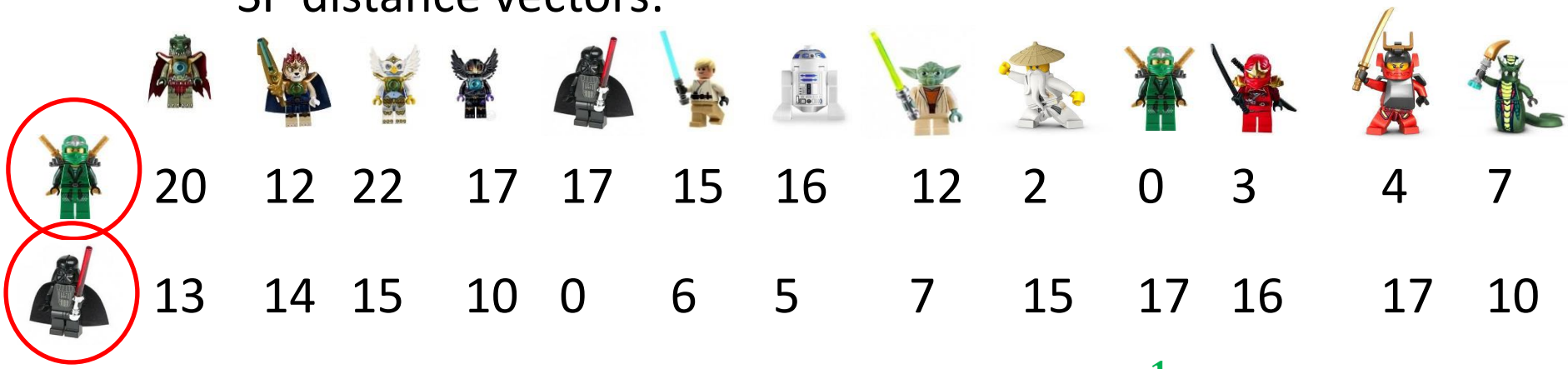
Specify how importance/relevance decays with distance

$$\alpha(d) = \begin{cases} \exp(-\lambda d) \text{ (exponential)}, \\ \frac{1}{d} \text{ (geomertic)}, \\ exp(-\lambda d^2) \text{ (Gaussian)} \\ d < \infty ?\ 1 : 0 \text{ (reachability)} \\ d < T ?\ 1 : 0 \text{ (threshold)} \end{cases}$$

- Similar to the role of kernel distribution in PPR

Closeness Vectors

SP distance vectors:



20	12	22	17	17	15	16	12	2	0	3	4	7
13	14	15	10	0	6	5	7	15	17	16	17	10

Closeness vectors:

$$\alpha(d_{ij}) = \frac{1}{1+d_{ij}} \qquad \beta \equiv 1$$

$\frac{1}{21}$	$\frac{1}{13}$	$\frac{1}{23}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{16}$	$\frac{1}{17}$	$\frac{1}{13}$	$\frac{1}{3}$	1	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{8}$
$\frac{1}{14}$	$\frac{1}{15}$	$\frac{1}{16}$	$\frac{1}{11}$	1	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{18}$	$\frac{1}{17}$	$\frac{1}{18}$	$\frac{1}{11}$

Closeness Centrality

- Classic (farness penalty)

[Bavelas/Sabidussi/Beuchamp 1950, 1965, 1966]

$$\boldsymbol{C(i) = (n-1) / \sum_{j} d_{ij}\, \beta(j)}$$

- Distance-decay (closeness reward)

[C' Kaplan 2004, Opsahl et. al. 2010, Dangalchev 2006, Boldi Vigna 2013,…]

$$\boldsymbol{C(i) = \sum_{j} \alpha(d_{ij})\, \beta(j)}$$

Closeness Matrix
(distance decay+topic filter)

$$\boldsymbol{c_{ij} = \alpha(d_{ij})\,\beta(j)}$$

$c_{i\cdot}$ (row $\boldsymbol{i}$) is the ***closeness vector*** of node $\boldsymbol{i}$

<u>Centrality:</u>

$$\boldsymbol{C(i) = \sum_{j} c_{ij}}$$

<u>Influence:</u>
of a set of nodes

$$\mathbf{Inf}(\boldsymbol{S}) = \sum_{j} \max_{i \in S} \boldsymbol{c_{ij}}$$

Example: (dist-decay) Closeness Centrality

SP distance vectors:

20	12	22	17	17	15	16	12	2	0	3	4	7

13	14	15	10	0	6	5	7	15	17	16	17	10

$$\alpha(d_{ij}) = \frac{1}{1+d_{ij}} \qquad \beta \equiv 1$$

Closeness vectors:

$\frac{1}{21}$	$\frac{1}{13}$	$\frac{1}{23}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{16}$	$\frac{1}{17}$	$\frac{1}{13}$	$\frac{1}{3}$	1	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{8}$

$\frac{1}{14}$	$\frac{1}{15}$	$\frac{1}{16}$	$\frac{1}{11}$	1	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{18}$	$\frac{1}{17}$	$\frac{1}{18}$	$\frac{1}{11}$

$C($ $) \approx 2.05$

$C($ $) \approx 2.11$

Example + topic filter: Closeness to Evil

0.8 0.1 0.1 0.8 1 0.1 0 0 0 0 0.1 0 0.9

$$\alpha(d_{ij}) = \frac{1}{1+d_{ij}} \qquad \beta \equiv evilness$$

$\frac{0.8}{21}$ $\frac{0.1}{13}$ $\frac{0.1}{23}$ $\frac{0.8}{18}$ $\frac{1}{18}$ $\frac{0.1}{16}$ $\frac{0}{17}$ $\frac{0}{13}$ $\frac{0}{3}$ 0 $\frac{0.1}{4}$ $\frac{0}{5}$ $\frac{0.9}{8}$

$\frac{0.8}{14}$ $\frac{0.1}{15}$ $\frac{0.1}{16}$ $\frac{0.8}{11}$ 1 $\frac{0.1}{7}$ $\frac{0}{6}$ $\frac{0}{8}$ $\frac{0}{16}$ $\frac{0}{18}$ $\frac{0.1}{17}$ $\frac{0}{18}$ $\frac{0.9}{11}$

$C($ $) \approx 2.05$ "total"

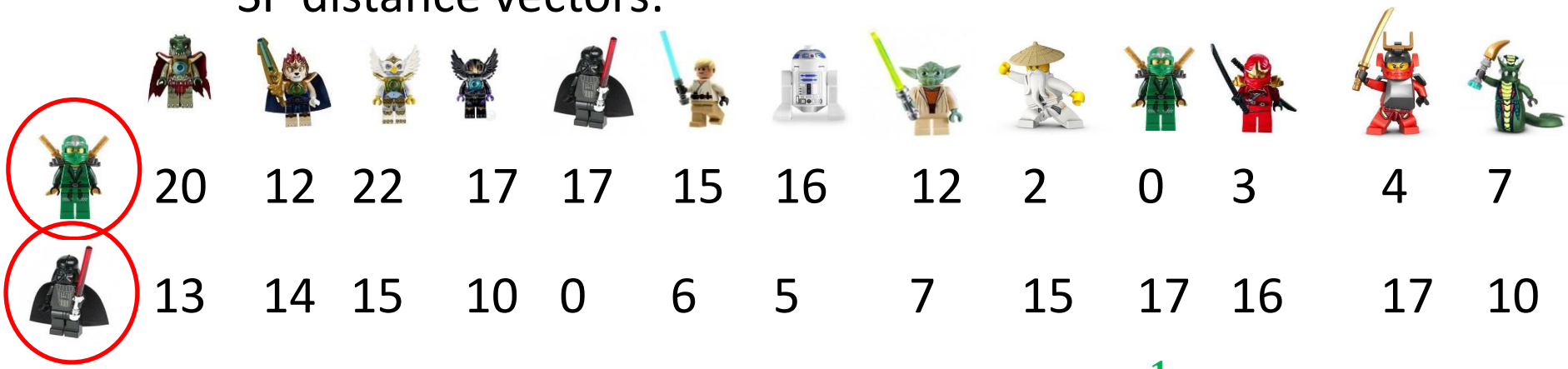
≈ 1.24 "evil part"

$C($ $) \approx 2.11$ "total"

≈ 0.29 "evil part"

Example: Influence from Vectors

SP distance vectors:



20	12	22	17	17	15	16	12	2	0	3	4	7
13	14	15	10	0	6	5	7	15	17	16	17	10

Closeness vectors:

$$\alpha(d_{ij}) = \frac{1}{1+d_{ij}} \qquad \beta \equiv 1$$

$\frac{1}{21}$	$\frac{1}{13}$	$\frac{1}{23}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{16}$	$\frac{1}{17}$	$\frac{1}{13}$	$\frac{1}{3}$	1	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{8}$
$\frac{1}{14}$	$\frac{1}{15}$	$\frac{1}{16}$	$\frac{1}{11}$	1	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{18}$	$\frac{1}{17}$	$\frac{1}{18}$	$\frac{1}{11}$

Inf(,) ≈ 3.64

$C($ $) \approx 2.11$

$C($ $) \approx 2.05$

Sum of max (Submodular)

Closeness Similarity: Local Measures

Local: $\mathbf{sim}(i, j)$ depends only on immediate neighbors of i, j

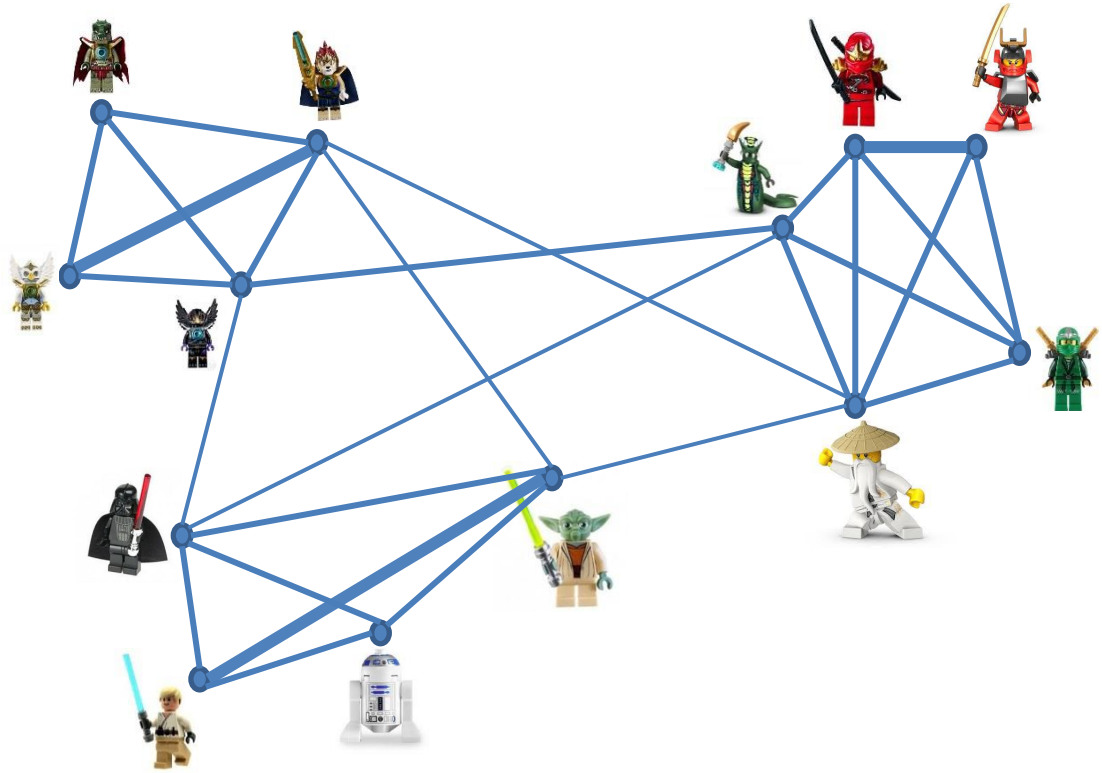
Jaccard: $\frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|}$ **Adamic Adar**: $\sum_{x \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log |\Gamma(x)|}$

J(,)= $\frac{2}{9}$

J(,)= $\frac{3}{5}$

AA(,)= $\frac{1}{\log 5}$

AA(,)= $\frac{3}{\log 4}$



Work well for $d_{ij} \leq 2$, but **no recall for** $d_{ij} > 2$

Closeness Similarity: Global Measures Based on full closeness vectors

$$c_{ij} = \alpha(d_{ij})\,\beta(j)$$

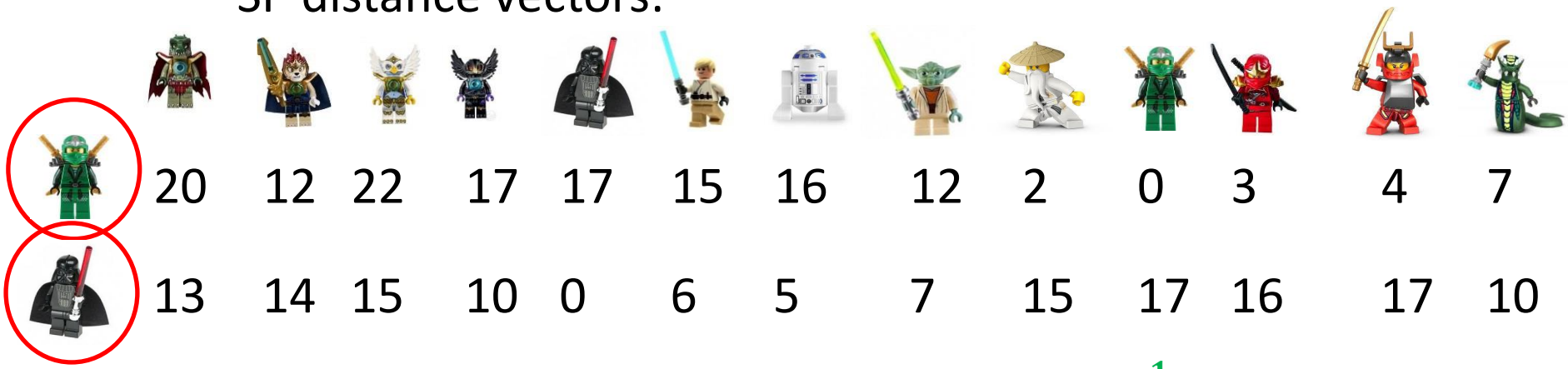
weighted Jaccard: $\mathbf{sim}(i,j) = \dfrac{\sum_h \beta(h)\,\min c_{ih}, c_{jh}}{\sum_h \beta(h)\max c_{ih}, c_{jh}}$

Cosine: $\mathbf{sim}(i,j) = c_{i\cdot} \cdot c_{j\cdot}$

L_p Distance: $L_p\text{-}\mathbf{dist}(i,j) = \left\| c_{i\cdot} - c_{j\cdot} \right\|_p$

Example: (global) Closeness Similarity

SP distance vectors:



20	12	22	17	17	15	16	12	2	0	3	4	7
13	14	15	10	0	6	5	7	15	17	16	17	10

Closeness vectors:

$$\alpha(d_{ij}) = \frac{1}{1+d_{ij}} \qquad \beta \equiv 1$$

$\frac{1}{21}$	$\frac{1}{13}$	$\frac{1}{23}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{16}$	$\frac{1}{17}$	$\frac{1}{13}$	$\frac{1}{3}$	1	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{8}$
$\frac{1}{14}$	$\frac{1}{15}$	$\frac{1}{16}$	$\frac{1}{11}$	1	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{18}$	$\frac{1}{17}$	$\frac{1}{18}$	$\frac{1}{11}$

Weighted Jaccard:

$$\frac{\sum_x \min C_{ix}, C_{jx}}{\sum_x \max C_{ix}, C_{jx}} \approx 0.21$$

Cosine similarity:

$$\frac{\sum_x C_{ix}\ C_{jx}}{||C_i||_2 ||C_j||_2} \approx 0.18$$

L_1 Distance:

$$\sum_x |C_{ix} - C_{jx}| \approx 2.8$$

So far: Definitions and some motivation of distance-based measures

Next: Scalability thru Sketches and estimators

Scalable Closeness Centrality

- Closeness centrality of a node v can be computed exactly using a SSSP computation from v.
- But this takes near-linear time per node: Does not scale when we want centralities of many or all nodes

We want to estimate *all* centralities within a small *relative error* using *near-linear computation*

- Classic (farness penalty): [C'DPW : COSN 2014]

  Previously additive error: [EW SODA 2001, OCL 2008]

- Distance-decay (closeness reward): All-distances sketches [C' 94], [C' Kaplan SIGMOD 2004] [C' : PODS 2014].

Scaling Up (closeness reward)

General Tools

- All-distances and reachability node sketches (ADS) [C' 94], [C' Kaplan 2004 [C' : PODS 2014].
- Estimators applicable with ADS: [C'K PODC 2007, VLDB 2008, SIGMETRICS 2008], [C' PODC 2014,KDD 2014]

Applications

- Distance-decay closeness centrality [C' 94], [C' Kaplan 2004] [Boldi Vigna 2013] [C' : PODS 2014].
- Closeness similarity [CDFGGW: COSN 2013]
- Influence computation and maximization [CWY KDD 2009] [C' Delling Pajor Werneck: CIKM 2014] [Du Song Gomez-Rodruiguez Zha NIPS 2013] [C' Delling Pajor Werneck 2015]

All-Distances Sketches (ADS) [C' 94]+

per-node summary structures of Un/Directed, Un/Weighted networks

For a node $i \in [n]$: $\mathrm{ADS}(i)$ "samples" the distance vector of i

$\boldsymbol{m}$ edges, $\boldsymbol{n}$ nodes, parameter $\boldsymbol{k} \geq 1$
trades-off sketch size and information

All-Distances Sketches: Definition

$\mathbf{ADS}(v)$ is a list of pairs of the form (i, d_{vi})

- Draw a random permutation of the nodes: $r : [n] \rightarrow [n]$
- $i \in \mathbf{ADS}(v) \Leftrightarrow r(i) <$ k^{th} smallest rank amongst nodes that are closer to v than i

This is a ***bottom-k ADS***, there are other ADS "flavors"

All-Distances Sketches: Definition

$\mathbf{ADS}(v)$ is a list of pairs of the form (i, d_{vi})

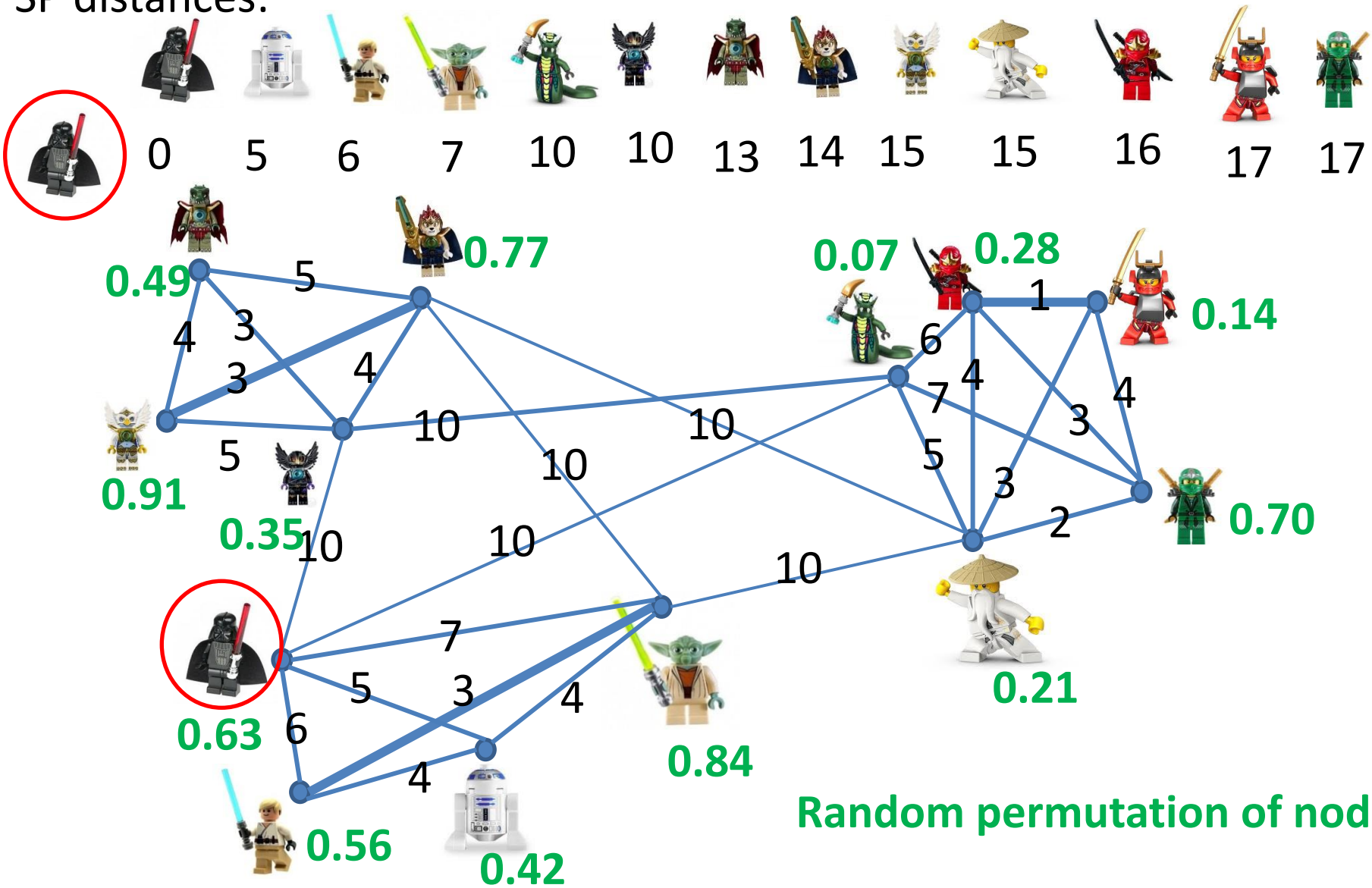
- Draw a random permutation of the nodes: $r : [n] \rightarrow [n]$
- $i \in \mathbf{ADS}(v) \iff r(i) <$ k^{th} smallest rank amongst nodes that are closer to v than i

Expected size of $\mathbf{ADS}(i)$ is $\sum_{j=1}^{n} \min\left\{1, \frac{k}{j}\right\} < k \ln n$

- The jth closest node to i is included with probability $\min\{1, \frac{k}{j}\}$

ADS example

SP distances:



ADS example $k = 1$

All nodes sorted by SP distance from

0.63 0.42 0.56 0.84 0.07 0.35 0.49 0.77 0.91 0.21 0.28 0.14 0.70

$k = 1$:

0.63 0.42 0.07

ADS example $k = 2$

Sorted by SP distances from

0.63 0.42 0.56 0.84 0.07 0.35 0.49 0.77 0.91 0.21 0.28 0.14 0.70

$k = 2$:

0.63 0.42 0.56 0.07 0.35 0.21 0.14

Sketch Coordination

We use the same permutation to obtain the ADS of all nodes, as a result:

- ADS Sketches of different nodes are *coordinated*: related in a way that is useful for queries that involve multiple nodes (similarities, influence, distance) [Brewer, Early, Joyce 1972]
- Generalize MinHash sketches by adding a time/distance dimension

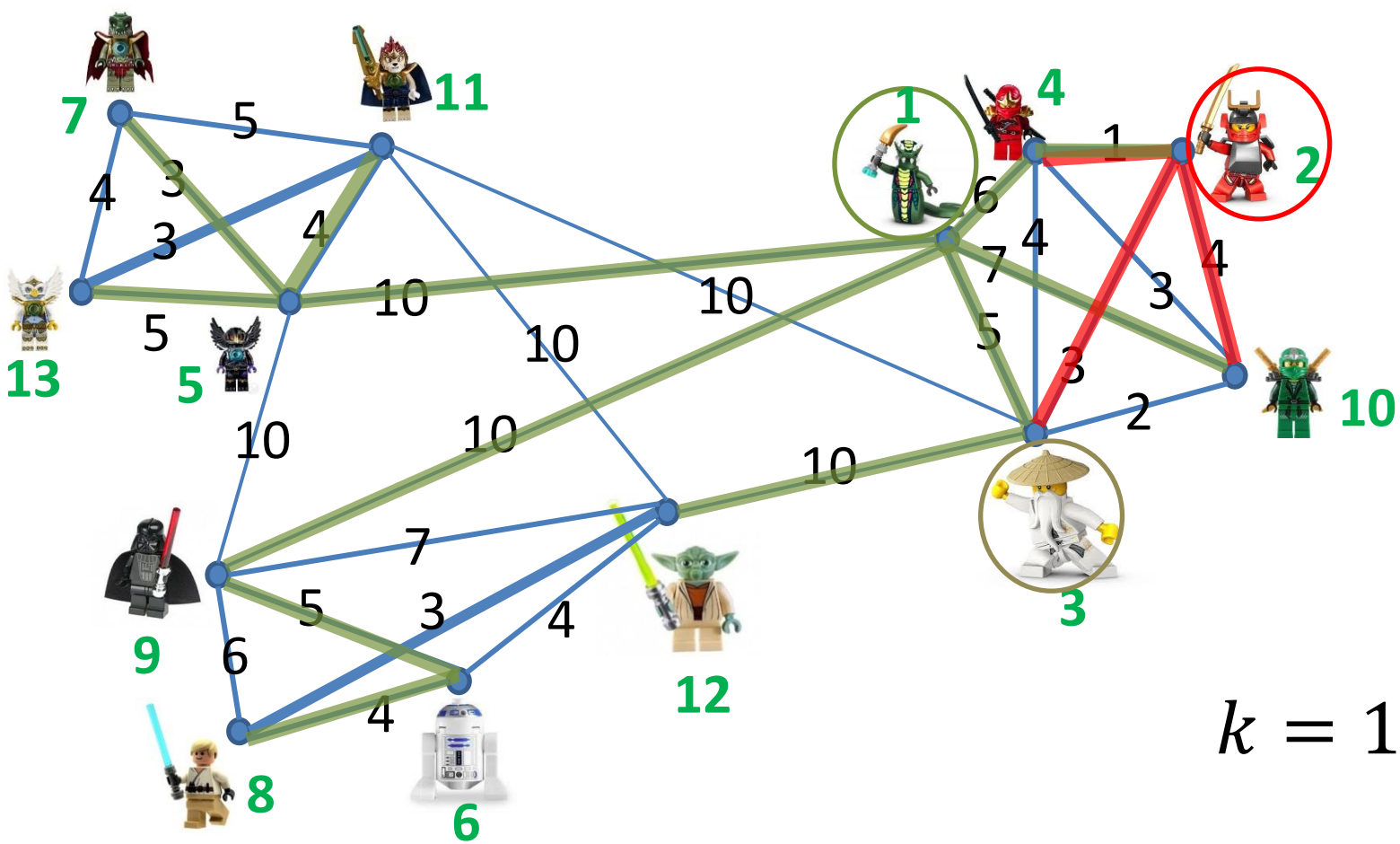
Computing ADSs Efficiently

- Pruned Dijkstra's Algorithm: Build ADS entries by **increasing permutation rank**.
- Local (Dynamic programming) Algorithm: Build ADS entries by **increasing distance.**

Either way, we scan outgoing edges of a node only after its ADS is modified ⇒ number of edge traversals is bounded by ADS size: $O(\boldsymbol{km} \ln \boldsymbol{n})$

Computing ADSs Efficiently

Perform pruned Dijkstra from nodes by increasing permutation rank:



$k = 1$

Estimation from sketches

Estimation with All-Distances Sketches

- Closeness centrality of node $\mathbf{i}$, from $\mathbf{ADS}(\boldsymbol{i})$ with NRMSE $\boldsymbol{O}(\frac{\mathbf{1}}{\sqrt{\boldsymbol{k}}})$ **(+concentration)** [C' 94, C' Kaplan 04, C' 14]
- Influence of $\boldsymbol{S}$, from $\mathbf{ADS}(\boldsymbol{i})$ $\mathbf{i} \in \boldsymbol{S}$, with NRMSE $\boldsymbol{O}(\frac{\mathbf{1}}{\sqrt{\boldsymbol{k}}})$ [C' 94, Chen WY KDD 09, Du SGZ NIPS 13, C' DPW 2014]+++
- Closeness similarity $\mathbf{sim}(\boldsymbol{i},\boldsymbol{j})$ (W/Jaccard , cosine), from $\mathbf{ADS}(\boldsymbol{i})$ and $\mathbf{ADS}(\boldsymbol{j})$ w. RMSE $\boldsymbol{O}(\frac{\mathbf{1}}{\sqrt{\boldsymbol{k}}})$ [C' DFGGW COSN 13]
- $\mathrm{L_p}$ **distance,** from $\mathbf{ADS}(\boldsymbol{i})$ and $\mathbf{ADS}(\boldsymbol{j})$ with RMSE $\boldsymbol{O}(\frac{\mathbf{1}}{\sqrt{\boldsymbol{k}}})$ **times max norm** using [C' KDD 2014]

Side note: ADSs can also be used as distance oracles (estimate pairwise distances), spanners, and more

Historic Inverse Probability (HIP) probability & estimator [C' 2014]

- The ***HIP inclusion probability*** p_{vi} of $i \in \mathrm{ADS}(v)$ is *conditioned* on fixed rank values of other nodes.
- To estimate centrality of v , we estimate the "**presence" of i with respect to v**: $I_{v \rightsquigarrow i}$ (=1 if $v \rightsquigarrow i$, 0 otherwise) using "inverse-probability" [HT52] $\boldsymbol{a_{vi} = \frac{1}{p}}$. This is unbiased (when $p > 0$).
- For bottom-k *ADS and* $r \sim U[0,1]$

$$p_{vi} = \mathrm{k^{th}}\{r(u) | u \in \mathrm{ADS}(v) \wedge d_{vu} < d_{vi}\}$$

Example: HIP estimates

Bottom-2 ADS of

	0.63	0.42	0.56	0.07	0.35	0.21	0.14
p:	1	1	0.63	0.56	0.42	0.35	0.21
$a = \frac{1}{p}$:	1	1	1.59	1.79	2.38	2.86	4.76

p:2nd smallest r value among closer nodes

HIP cardinality estimate

Bottom-2 ADS of

distance:	0	5	6	10	10	15	17
a:	1	1	1.59	1.79	2.38	2.86	4.76

Query: $\boldsymbol{n_6(v) = |\{\, i \,|d_{vi} \leq 6 \,\}| = \sum_i I_{d_{vi} \leq 6}}$

$$\widehat{\boldsymbol{n_6(v)}} = \sum_{(i,d_{vi}) \in ADS(v) \,|\, d_{vi} \leq 6} \boldsymbol{a_{vi}} = 1 + 1 + 1.59 = 3.59$$

Quality of HIP cardinality Estimate

<u>Lemma:</u> The HIP neighborhood cardinality estimator

$$\widehat{n_d(v)} = \sum_{(i,d_{vi}) \in ADS(v) \mid d_{vi} \le d} a_{vi}$$

Has Coefficient of Variation $\frac{\sigma}{\mu} \le \frac{1}{\sqrt{2k-2}}$

This is $\sqrt{2}$ improvement over "MinHash" estimators [C' 94]. Also optimally uses sketch.

HIP estimates of Centrality

$$C_v = \sum_i \alpha(d_{vi})\beta(i)$$

- α non increasing; β some filter

$$\widehat{C_v} = \sum_{i \in ADS(v)} a_{vi}\, \alpha(d_{vi})\, \beta(i)$$

HIP estimates: closeness to good/evil

Bottom-2 ADS of

distance:	0	5	6	10	10	15	17
a:	1	1	1.59	1.79	2.38	2.86	4.76
$\boldsymbol{\beta}$:	**0**	**1**	**1**	**0.2**	**0.1**	**1**	**0.9**

Filter: $\beta \in [0,1]$ measures "goodness"

Distance-decay "kernel": $\boldsymbol{e^{-d_{vi}}}$

$$\widehat{C_v} = \sum_{i \in ADS(v)} a_{vi}\beta(i)e^{-d_{vi}} = e^{-5} + 1.59e^{-6} + 0.2e^{-10} + \cdots$$

Similarity/Influence Estimation

We work with HIP inclusions and sum estimators (estimate separately contribution of each node)

We use *monotone estimation formulation* [C'K Random 2014; C' PODC 2014] and the *L* estimator* (unique optimal monotone unbiased nonnegative sum estimator)

- Influence: can be estimated by merging ADS sketches and applying centrality estimator, but L* is tighter
- Similarity: inverse-probability on joint inclusion may not even apply, L* gets around the problem

Estimating Jaccard Closeness Similarity

SP-distance; using kernel $\alpha(d_{ij}) = \frac{1}{1+d_{ij}}$:

$$\mathbf{sim}(i,j) = \frac{\sum_x \min C_{ix}, C_{jx}}{\sum_x \max C_{ix}, C_{jx}} = \frac{\sum_x \min \frac{1}{d_{ix}+1}, \frac{1}{d_{jx}+1}}{\sum_x \max \frac{1}{d_{ix}+1}, \frac{1}{d_{jx}+1}}$$

- Sum estimator: For all x obtain unbiased estimates

$\widehat{N_x}$ **of min** $\frac{1}{d_{ix}+1}, \frac{1}{d_{jx}+1}$; $\widehat{X_x}$ **of max** $\frac{1}{d_{ix}+1}, \frac{1}{d_{jx}+1}$;

$$\widehat{sim}(i,j) = \frac{\sum_x \widehat{N_x}}{\sum_x \widehat{X_x}}$$

- $\widehat{X_x} > 0$ or $\widehat{N_x} > 0$ **only** when $x \in ADS(i) \cup ADS(j)$
  $\Longrightarrow$ easy to compute $\widehat{sim}(i,j)$ using only $ADS(i) \cup ADS(j)$
- We use L* for X_x. We show $\widehat{N_x}$ (inverse probability). Get relative error $\frac{1}{\sqrt{k}}$ on denominator, additive error of same order on numerator.

Estimating Jaccard Closeness Similarity

Estimate $\widehat{N_x}$ for $\min \frac{1}{d_{ix}+1}, \frac{1}{d_{jx}+1}$:

For $x \in ADS(i) \cap ADS(j)$

- We compute HIP probability p for inclusion in intersection
- An inverse probability estimate:

$$\widehat{N_x} = \frac{1}{p} \min \frac{1}{d_{ix}+1}, \frac{1}{d_{jx}+1}$$

Otherwise, estimate is $\widehat{N_x} = 0$

➢ This is the best we can do because if x is not in intersection, data is consistent with 0 value ⇒ any nonzero unbiased estimator must return 0

Estimating Jaccard Closeness Similarity

Estimate $\widehat{X_x}$ for $\mathbf{max}\,\frac{1}{d_{ix}+1}, \frac{1}{d_{jx}+1}$

For $x \in ADS(i) \cup ADS(j)$

- We can never determine X_x if x is only reachable from one of i, j. We can only lower bound X_x.
- The L* estimator [C' 2014] is an optimal extension of inverse probability to such a setting.

NN-rank based closeness similarity

$$\mathbf{sim}(i,j) = \frac{\sum_x \min \frac{1}{\pi_{ix}}, \frac{1}{\pi_{jx}}}{\sum_x \max \frac{1}{\pi_{ix}}, \frac{1}{\pi_{jx}}}$$

π_{ix}: Dijkstra (NN) rank of x with respect to i

<u>Lemma:</u>

$\mathbf{sim}(i,j)$ is approximated well by $\frac{|ADS(i) \cap ADS(j)|}{|ADS(i) \cup ADS(j)|}$

"proof:"

- $\Pr[\, x \in ADS(i) \cap ADS(j)] \approx k \min \frac{1}{\pi_{ix}}, \frac{1}{\pi_{jx}}$
- $\Pr[\, x \in ADS(i) \cup ADS(j)] \approx k \max \frac{1}{\pi_{ix}}, \frac{1}{\pi_{jx}}$

⇒ ADSs can be stored very compactly: no distances, hash of nodes

Next:
Enhancing distance/reachability-based models with REL

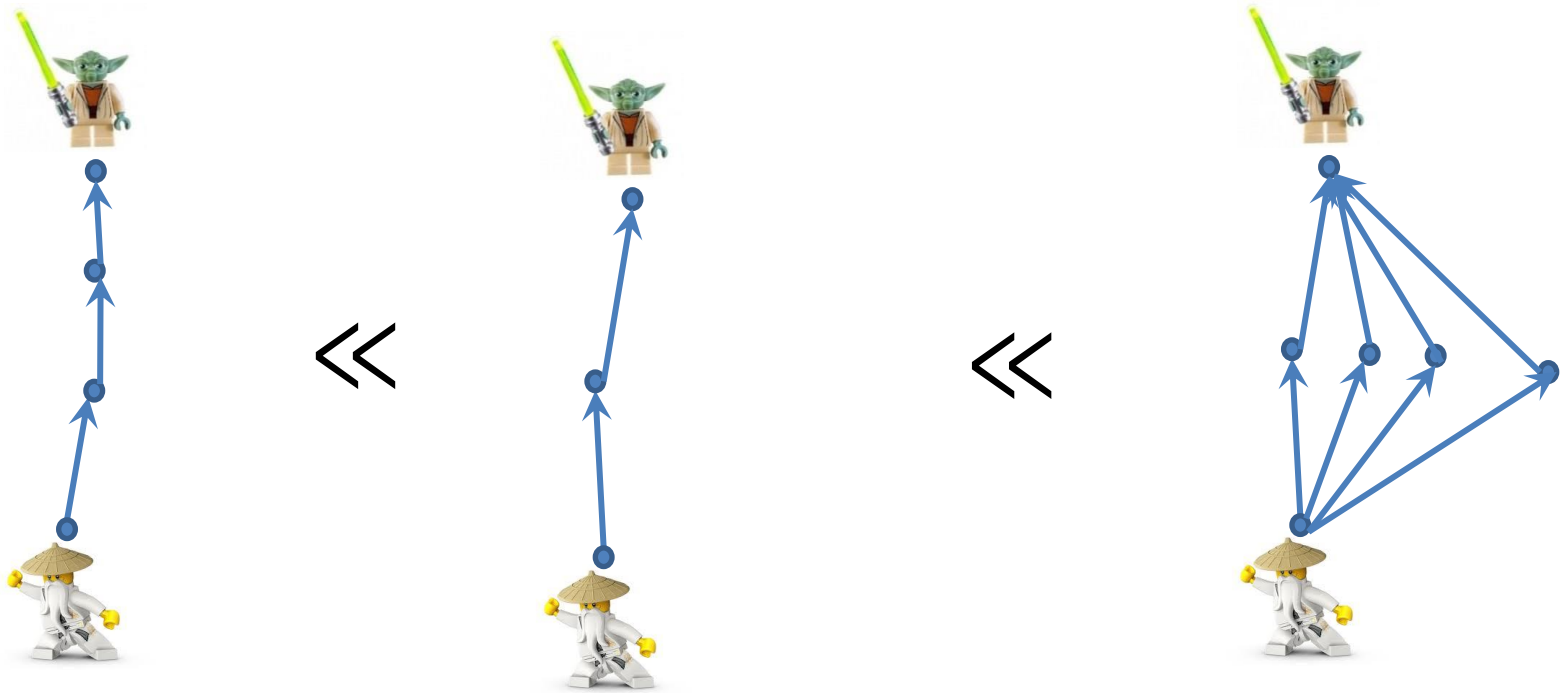
REL: randomizing lengths/presence of edges

- (Implicit) The Independent Cascade Diffusion model [Kempe Klienberg Tardos KDD 2003]
- Closeness similarity [C'DFGGW 2013]
- Timed Diffusion [Gomez-Rodriguez BS 2011, Du SGZ 2013, C'DPW 2014,….]

Strength of relation between two entities

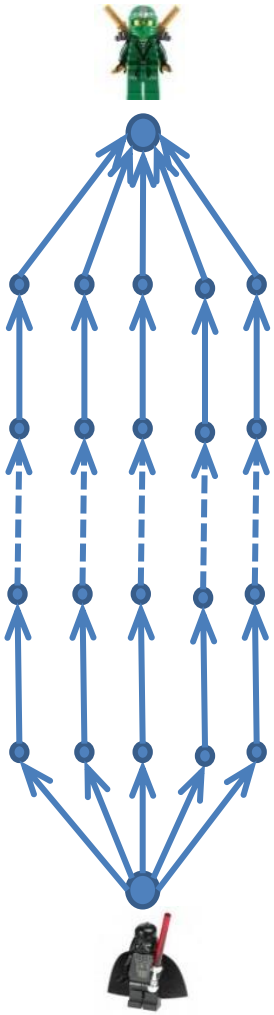
Basic Intuitions for dependence on path ensemble:

- Increase with shorter paths
- Increase with more (redundant paths)



Boosting distance by Randomizing Edge Lengths (REL)

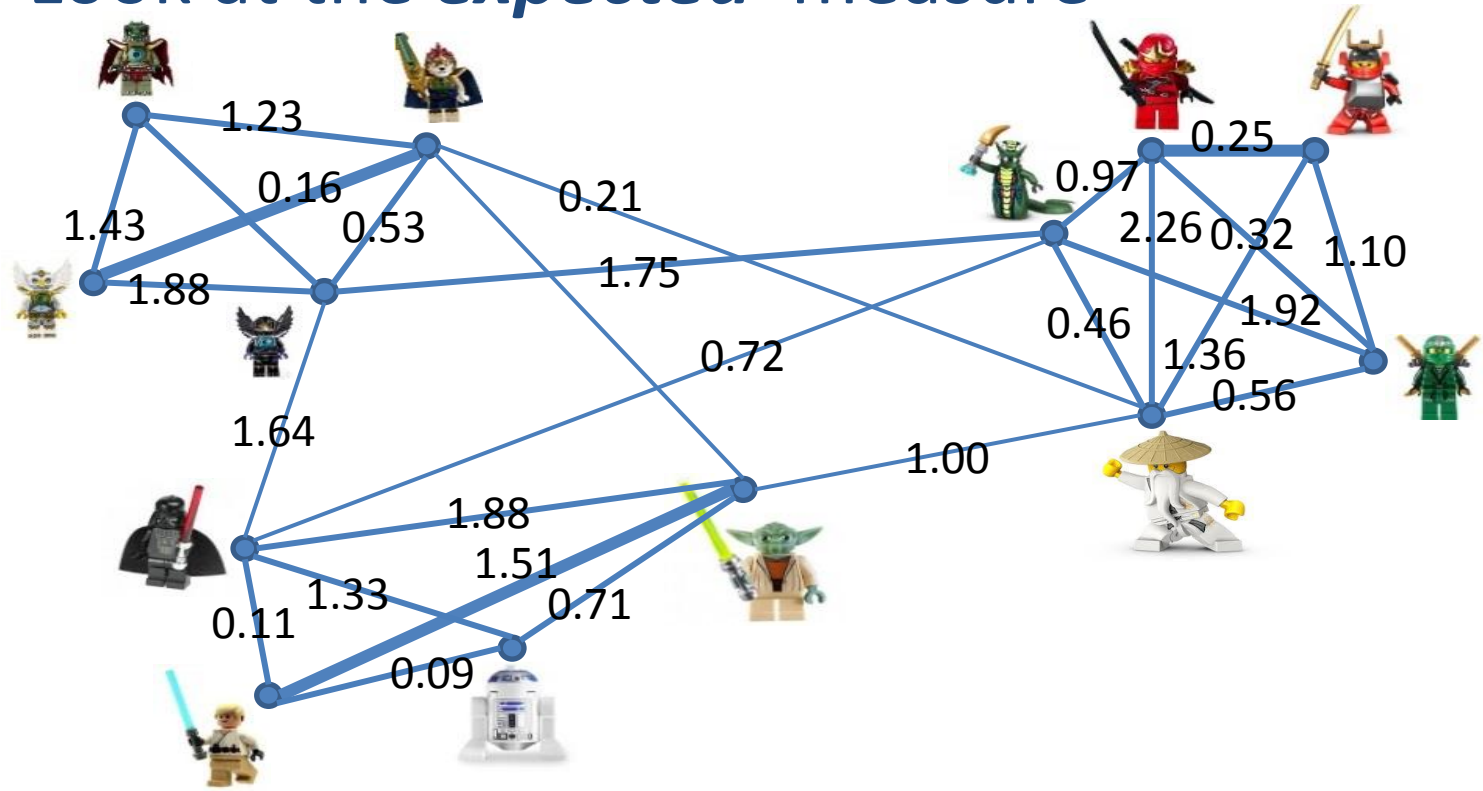
We expect the strength of a relation to
- ***Increase*** with the multiplicity of paths
- ***Decrease*** with the length of paths



- "Eigenvector" measures:
  Rooted PageRank, **RWR**, **Hitting Time**, **Commute Time**, **Eff. Resistance**, **Katz**
  Satisfy this *strictly*.
- **SP distances**: *weakly*

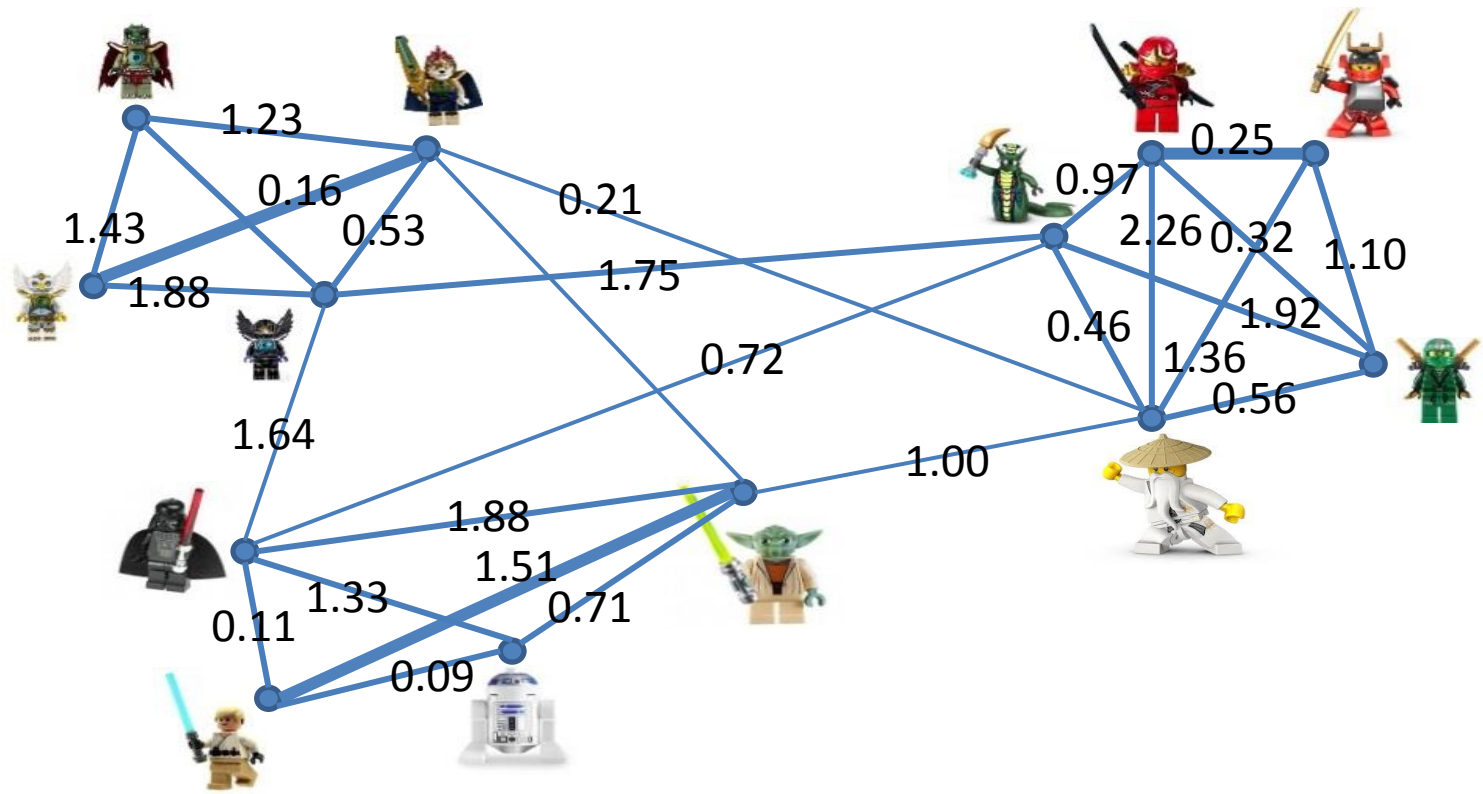
Randomizing Edge Lengths (REL)

- Replace edge lengths (or presence) with independent random variables
- Look at the ***expected*** measure



Randomizing Edge Lengths (REL)

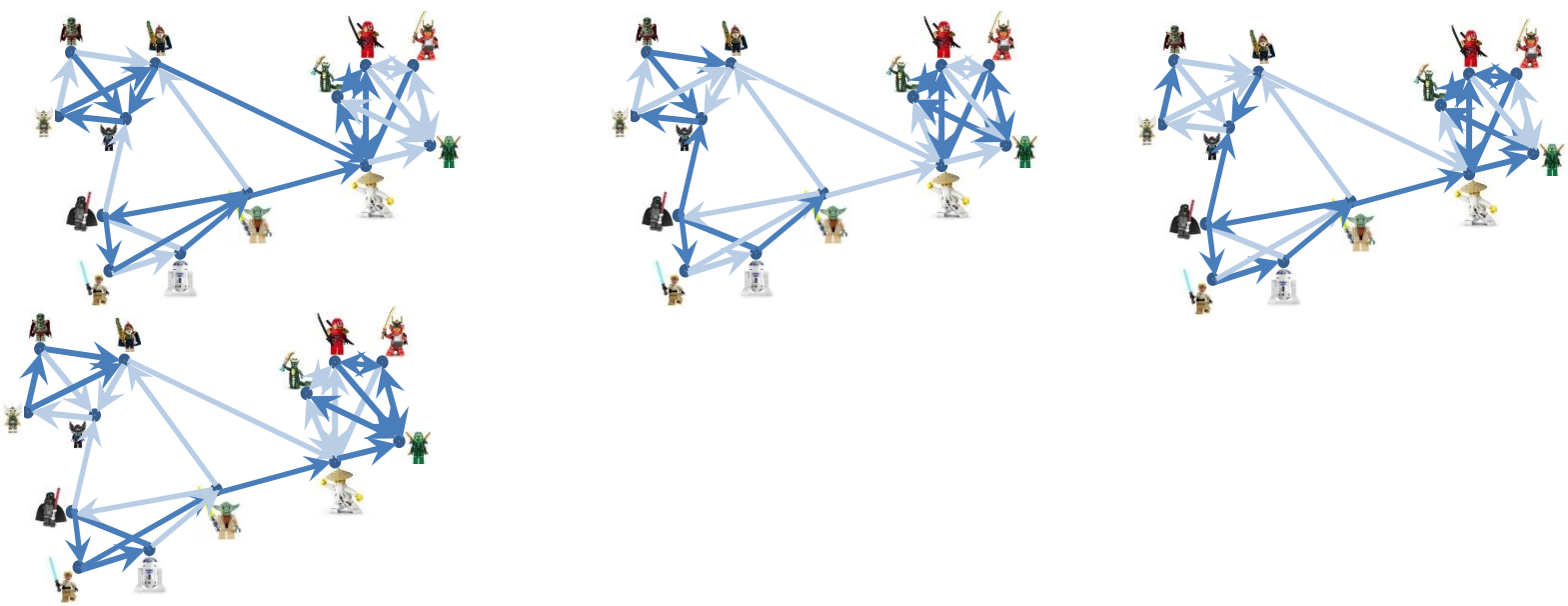
Expected distance is lower with multiple paths: expectation of the minimum < minimum of expectations



Randomizing Edge Lengths (REL)

Scalability:

- use Monte Carlo simulations of model to generate fixed instances (graphs)
- Build sketches for multiple instances



Benefits of REL in network analysis

- Measure does reward for paths multiplicity
- Robustness: Measure is not sensitive to small changes in edge weights (edge presence for reachability)

Next:
Some (preliminary) experimental results

Scalability: Timed Influence with REL

Combined ADS sketches for all nodes, obtained from $\ell = 64$ instances generated by assigning independent Exp distributed edge lengths. Sketch parameter $k = 64$

Centrality (1 seed) and Influence (50 seeds) queries

				Queries $\alpha(x) = exp(-10x)$			
			preproc	1 seed		50 seeds	
	#nodes	#edges	[h:m]	time μs	%err	time μs	%err
Slashdot	77K	828K	1:10	46	5.6%	11880	0.4%
Gowalla	197K	1.9M	3:55	52	3.8%	17100	0.4%
Twitter F	456K	15M	19:33	51	3.8%	13800	0.7%

[C' Delling Pajor Werneck 2014]

Closeness Similarity evaluation

[C' Delling Fuchs Goldberg Goldszmid Werneck COSN 2013]

Data Sets

	Nodes x 10^6	Edges x 10^6	ADS label size
arXiv	0.4	28.7	37.9
DBLP	1.1	9.2	39.1
twitter	29.6	603.9	101.6
smallworld	1.0	6.0	40.7

Similarity Evaluation

Spearman coefficient: correlation between meta-data ranking and similarity measure ranking(1 = perfect match, 0= random rankings).

On pairs selected “uniformly” by “ground-truth” similarity.

	arXiv	DBLP	Twitter	SmallWord
Adamic-Adar	0.626	0.746	0.548	0.000
hops	**0.752**	0.748	0.169	0.767
SP distance	0.590	0.634	0.140	0.767
RWR 0.75		0.734		0.286
RWR 0.50		0.737		0.617
RWR 0.25		0.740		0.791
RWR 0.00		0.500		**0.915**
Closeness	0.641	0.742	0.613	0.609
Closeness REL	0.634	**0.752**	**0.649**	0.808

Similarity Evaluation

- No clear winner (networks too different)
- Local measures are limited
- RWR has very good recall (with tuning) but computationally expensive
- Closeness (+REL) : robust, good recall, fast queries (microseconds)

	arXiv	DBLP	Twitter	SmallWord
Adamic-Adar	0.626	0.746	0.548	0.000
hops	**0.752**	0.748	0.169	0.767
SP distance	0.590	0.634	0.140	0.767
RWR 0.75		0.734		0.286
RWR 0.50		0.737		0.617
RWR 0.25		0.740		0.791
RWR 0.00		0.500		**0.915**
Closeness	0.641	0.742	0.613	0.609
Closeness REL	0.634	**0.752**	**0.649**	0.808

Conclusion

Distance/reachability based measures of centrality, influence, and similarity: global measures that are flexible and highly scalable

- Presented a unified treatment
- Scalability through:
  - All-Distances and reachability sketches
  - Estimators applicable to sketches
- Future: Model fitting framework, Evaluations, Algorithms Engineering

Thank you!

Components of my work are joint work with (subsets of):

Daniel Delling, Fabian Fuchs, Andrew Goldberg, Moises Goldszmidt, Haim Kaplan, Thomas Pajor, and Renato Werneck