Graph distance distribution for social network mining

Paolo Boldi, Marco Rosa, Sebastiano Vigna Laboratory for Web Algorithmics Università degli Studi di Milano, Italy

> Lars Backstrom, Johan Ugander Facebook

• Computing distances in large graphs (using HyperANF)

- *Computing distances* in large graphs (using HyperANF)
- Running HyperANF on *Facebook* (the largest Milgram-like experiment ever performed)

- *Computing distances* in large graphs (using HyperANF)
- Running HyperANF on *Facebook* (the largest Milgram-like experiment ever performed)
- Other uses of distances (in particular: robustness)

Prelude Milgram's experiment is 45

• M. Kochen, I. de Sola Pool: *Contacts and influences*. (Manuscript, early 50s)

- M. Kochen, I. de Sola Pool: *Contacts and influences*. (Manuscript, early 50s)
- A. Rapoport, W.J. Horvath: *A study of a large sociogram*. (Behav.Sci. 1961)

- M. Kochen, I. de Sola Pool: *Contacts and influences*. (Manuscript, early 50s)
- A. Rapoport, W.J. Horvath: *A study of a large sociogram*. (Behav.Sci. 1961)
- S. Milgram, An experimental study of the small world problem. (Sociometry, 1969)

• 300 people (*starting population*) are asked to dispatch a parcel to a single individual (*target*)

- 300 people (*starting population*) are asked to dispatch a parcel to a single individual (*target*)
- The target was a Boston stockbroker

- 300 people (*starting population*) are asked to dispatch a parcel to a single individual (*target*)
- The target was a Boston stockbroker
- The starting population is selected as follows:

- 300 people (*starting population*) are asked to dispatch a parcel to a single individual (*target*)
- The target was a Boston stockbroker
- The starting population is selected as follows:
 - 100 were random Boston inhabitants (group A)

- 300 people (*starting population*) are asked to dispatch a parcel to a single individual (*target*)
- The target was a Boston stockbroker
- The starting population is selected as follows:
 - 100 were random Boston inhabitants (group A)
 - 100 were random Nebraska strockbrokers (group B)

- 300 people (*starting population*) are asked to dispatch a parcel to a single individual (*target*)
- The target was a Boston stockbroker
- The starting population is selected as follows:
 - 100 were random Boston inhabitants (group A)
 - 100 were random Nebraska strockbrokers (group B)
 - 100 were random Nebraska inhabitants (group C)

• Rules of the game:

• Rules of the game:

 parcels could be directly sent *only* to someone the sender knows personally

• Rules of the game:

- parcels could be directly sent *only* to someone the sender knows personally
- 453 intermediaries happened to be involved in the experiments (besides the starting population and the target)

• Questions Milgram wanted to answer:

• Questions Milgram wanted to answer:

• How many parcels will reach the target?

• Questions Milgram wanted to answer:

- How many parcels will reach the target?
- What is the distribution of the number of hops required to reach the target?

• Questions Milgram wanted to answer:

- How many parcels will reach the target?
- What is the distribution of the number of hops required to reach the target?
- Is this distribution different for the three starting subpopulations?

• Answers:

• Answers:

• How many parcels will reach the target? 29%

• Answers:

- How many parcels will reach the target? 29%
- What is the distribution of the number of hops required to reach the target? **Avg. was 5.2**

• Answers:

- How many parcels will reach the target? 29%
- What is the distribution of the number of hops required to reach the target? **Avg. was 5.2**
- Is this distribution different for the three starting subpopulations? Yes: avg. for groups A/B/C was 4.6/5.4/5.7, respectively

Chain lengths



• *Six degrees of separation* slipped away from the scientific niche to enter the world of popular immagination:

- *Six degrees of separation* slipped away from the scientific niche to enter the world of popular immagination:
 - "Six degrees of separation" is a play by John Guare...

- *Six degrees of separation* slipped away from the scientific niche to enter the world of popular immagination:
 - "Six degrees of separation" is a play by John Guare...
 - ...a movie by Fred Schepisi...
Milgram's popularity

- Six degrees of separation slipped away from the scientific niche to enter the world of popular immagination:
 - "Six degrees of separation" is a play by John Guare...
 - ...a movie by Fred Schepisi...
 - ...a song sung by dolls in their national costume at Disneyland in a heart-warming exhibition celebrating the connectedness of people all

 "Could it be a big world after all? (The sixdegrees-of-separation myth)" (Judith S. Kleinfeld, 2002)

- "Could it be a big world after all? (The sixdegrees-of-separation myth)" (Judith S. Kleinfeld, 2002)
 - The vast majority of chains were never completed

- "Could it be a big world after all? (The sixdegrees-of-separation myth)" (Judith S. Kleinfeld, 2002)
 - The vast majority of chains were never completed
 - Extremely difficult to reproduce

• But what did Milgram's experiment reveal, after all?

• But what did Milgram's experiment reveal, after all?

i) That the world is small

- But what did Milgram's experiment reveal, after all?
 - i) That the world is small
 - ii)That people are able to exploit this smallness

HyperBall A tool to compute distances in large graphs

• You want to study the properties of a *huge* graph (typically: a social network)

- You want to study the properties of a *huge* graph (typically: a social network)
- You want to obtain some information about its *global* structure (not simply triangle-counting/degree distribution/etc.)

- You want to study the properties of a *huge* graph (typically: a social network)
- You want to obtain some information about its *global* structure (not simply triangle-counting/degree distribution/etc.)
- A natural candidate: distance distribution

 Given a graph, d(x,y) is the length of the shortest path from x to y (∞ if one cannot go from x to y)

- Given a graph, d(x,y) is the length of the shortest path from x to y (∞ if one cannot go from x to y)
- For *undirected* graphs, d(x,y)=d(y,x)

- Given a graph, d(x,y) is the length of the shortest path from x to y (∞ if one cannot go from x to y)
- For *undirected* graphs, d(x,y)=d(y,x)
- For every *t*, count the number of pairs (*x*,*y*) such that d(x,y)=t

- Given a graph, d(x,y) is the length of the shortest path from x to y (∞ if one cannot go from x to y)
- For *undirected* graphs, d(x,y)=d(y,x)
- For every *t*, count the number of pairs (*x*,*y*) such that d(x,y)=t
- The fraction of pairs at distance *t* is (the density function of) a distribution

• How can one compute the distance distribution?

- How can one compute the distance distribution?
 - Weighted graphs: Dijkstra (single-source: O(n²)), Floyd-Warshall (all-pairs: O(n³))

- How can one compute the distance distribution?
 - Weighted graphs: Dijkstra (single-source: O(n²)), Floyd-Warshall (all-pairs: O(n³))
 - In the unweighted case:

- How can one compute the distance distribution?
 - Weighted graphs: Dijkstra (single-source: O(n²)), Floyd-Warshall (all-pairs: O(n³))
 - In the unweighted case:
 - a single BFS solves the single-source version of the problem: O(m)

- How can one compute the distance distribution?
 - Weighted graphs: Dijkstra (single-source: O(n²)), Floyd-Warshall (all-pairs: O(n³))
 - In the unweighted case:
 - a single BFS solves the single-source version of the problem: O(m)
 - if we repeat it from every source: O(nm)

• Sample at random pairs of nodes (x,y)

- Sample at random pairs of nodes (x,y)
- Compute d(x,y) with a BFS from x

- Sample at random pairs of nodes (x,y)
- Compute d(x,y) with a BFS from x
- (Possibly: reject the pair if d(x,y) is infinite)

• For every *t*, the fraction of sampled pairs that were found at distance *t* are an estimator of the value of the probability mass function

- For every *t*, the fraction of sampled pairs that were found at distance *t* are an estimator of the value of the probability mass function
- Takes a BFS for every pair O(m)

• Sample at random a source x

- Sample at random a source x
- Compute a full BFS from x
• It is an unbiased estimator only for undirected and connected graphs

• It is an unbiased estimator only for undirected and connected graphs

• Uses anyway BFS...

- It is an unbiased estimator only for undirected and connected graphs
- Uses anyway BFS...
 - ...not cache friendly

- It is an unbiased estimator only for undirected and connected graphs
- Uses anyway BFS...
 - ...not cache friendly
 - ...not compression friendly

Cohen's sampling

Cohen's sampling

 Edith Cohen [JCSS 1997] came out with a very general framework for size estimation: powerful, but doesn't scale well, it is not easily parallelizable, requires direct access

• Basic idea: Palmer et. al, KDD '02

- Basic idea: Palmer et. al, KDD '02
- Let B_t(x) be the ball of radius t about x (the set of nodes at distance ≤t from x)

- Basic idea: Palmer et. al, KDD '02
- Let B_t(x) be the ball of radius t about x (the set of nodes at distance ≤t from x)
- Clearly $B_o(x) = \{x\}$

- Basic idea: Palmer et. al, KDD '02
- Let B_t(x) be the ball of radius t about x (the set of nodes at distance ≤t from x)
- Clearly $B_o(x) = \{x\}$
- Moreover $B_{t+1}(x) = \bigcup_{x \to y} B_t(y) \bigcup \{x\}$

- Basic idea: Palmer et. al, KDD '02
- Let B_t(x) be the ball of radius t about x (the set of nodes at distance ≤t from x)
- Clearly $B_o(x) = \{x\}$
- Moreover $B_{t+1}(x) = \bigcup_{x \to y} B_t(y) \bigcup \{x\}$
- So computing B_{t+1} starting from B_t one just need a single (sequential) scan of the graph
















































Another round...



Every set requires O(n) bits, hence O(n²) bits overall

- Every set requires O(n) bits, hence O(n²) bits overall
- Too many!

- Every set requires O(n) bits, hence O(n²) bits overall
- Too many!
- What about using approximated sets?

- Every set requires O(n) bits, hence O(n²) bits overall
- Too many!
- What about using approximated sets?
- We need *probabilistic counters*, with just two primitives: add and size?

- Every set requires O(n) bits, hence O(n²) bits overall
- Too many!
- What about using approximated sets?
- We need *probabilistic counters*, with just two primitives: add and size?
- Very small!

• We used HyperLogLog counters [Flajolet *et al.*, 2007]

- We used HyperLogLog counters [Flajolet *et al.*, 2007]
- With 40 bits you can count up to 4 billion with a standard deviation of 6%

- We used HyperLogLog counters [Flajolet *et al.*, 2007]
- With 40 bits you can count up to 4 billion with a standard deviation of 6%
- Remember: one set per node!

• Every single counter has a guaranteed *relative standard deviation* (depending only on the number of registers per counter)

- Every single counter has a guaranteed *relative standard deviation* (depending only on the number of registers per counter)
- This implies a guarantee on the *summation* of the counters

- Every single counter has a guaranteed *relative standard deviation* (depending only on the number of registers per counter)
- This implies a guarantee on the *summation* of the counters
- This gives in turn precision bounds on the estimated distribution with respect to the real one

• We use *broadword programming* to compute efficiently unions

- We use *broadword programming* to compute efficiently unions
- *Systolic computation* for on-demand updates of counters

- We use *broadword programming* to compute efficiently unions
- *Systolic computation* for on-demand updates of counters
- Exploited *microparallelization* of multicore architectures

Footprint

• Scalability: a minimum of 20 bytes per node

- On a 2TiB machine, 100 billion nodes
- Graph structure is accessed by memory-mapping in a compressed form (WebGraph)
- Pointer to the graph are store using succinct lists (Elias-Fano representation)

Performance

- On a 177K nodes / 2B arcs graph
- Hadoop: 2875s per iteration [Kang, Papadimitriou, Sun and H. Tong, 2011]
- HyperBall on this laptop: 70s per iteration
- On a 32-core workstation: 23s per iteration
- On ClueWebo9 (4.8G nodes, 8G arcs) on a 40-core workstation: 141m (avg. 40s per iteration)





• HyperBall is available within the webgraph package



- HyperBall is available within the webgraph package
- Download it from



- HyperBall is available within the webgraph package
- Download it from
 - <u>http://webgraph.di.unimi.it</u>/

Try it!

- HyperBall is available within the webgraph package
- Download it from
 - <u>http://webgraph.di.unimi.it</u>/
- Or google for webgraph

Running it on Facebook! [with Lars Backstrom and Johan Ugander]





• Facebook opened up to non-college students on September 26, 2006

Facebook

- Facebook opened up to non-college students on September 26, 2006
- So, between 1 Jan 2007 and 1 Jan 2008 the number of users exploded

Experiments (time)

• We ran our experiments on snapshots of facebook

• Jan 1, 2007

- Jan 1, 2008 ...
- Jan 1, 2011
- [current] May, 2011

Experiments (dataset)

• We considered:

• fb: the whole facebook

- it / se: only Italian / Swedish users
- it+se: only Italian & Swedish users
- us: only US users
- Based on users' current geo-IP location

Active users

- We only considered *active* users (users who have done some activity in the 28 days preceding 9 Jun 2011)
- So we are not considering "old" users that are not active any more
- For fb [current] we have about 750M nodes

Distance distribution (fb)

Distance distribution (fb)



distance

Distance distribution (fb)



distance


















it 2011

















Average distance





6

7 8

5 6

4

2007

2008

avg. distance



year

current





year

2010

2011



2009

	2008	curr
it	6.58	3.90
se	4.33	3.89
it+se	4.9	4.16
us	4.74	4.32
fb	5.28	4.74

Average distance





6

7 8

9

S

4

2007

2008

avg. distance



current



year

2010

2011



2009



Average distance











	2008	curr
it	6.58	3.90
se	4.33	3.89
it+se	4.9	4.16
us	4.74	4.32
fb	5.28	4.74

Effective diameter (@ 90%)

effective diameters

		2008	curr
it	t	9.0	5.2
S	e	5.9	5.3
it +s	te	6.8	5.8
u	s	6.5	5.8
fl)	7.0	6.2

Harmonic diameter

harmonic diameters

	2008	curr
it	23.7	3.4
se	4.5	4.0
it +se	5.8	3.8
us	4.6	4.4
fb	5.7	4.6

Average degree vs. density (fb)

	Avg. degree	Density
2009	88.7	6.4 * 10 ⁻⁷
2010	113.0	3.4 * 10 ⁻⁷
2011	169.0	3.0 * 10 ⁻⁷
CUTT	190.4	2.6 * 10 ⁻⁷

Actual diameter

	2008	curr
it	>29	=25
se	>16	=25
it+se	>21	=27
US	>17	=30
fb	>16	>58

Actual diameter

Used the fringe/double-sweep technique for "="

	2008	curr
it	>29	=25
se	>16	=25
it+se	>21	=27
US	>17	=30
fb	>16	>58

Other applications Spid, network robustness and more...

What are distances good for?

What are distances good for?

• Network models are usually studied on the base of the local statistics they produce

What are distances good for?

- Network models are usually studied on the base of the local statistics they produce
- Not difficult to obtain models that behave correctly locally (i.e., as far as degree distribution, assortativity, clustering coefficients... are concerned)

Global = more informative!

An application

An application

• An application: use the distance distribution as a graph *digest*

An application

- An application: use the distance distribution as a graph *digest*
- Typical example: if I modify the graph with a certain criterion, how much does the distance distribution change?

Node elimination
Node elimination

• Consider a certain ordering of the vertices of a graph

Node elimination

- Consider a certain ordering of the vertices of a graph
- Fix a threshold ϑ, delete all *vertices* (and all incident arcs) in the specified order, until ϑm arcs have been deleted

Node elimination

- Consider a certain ordering of the vertices of a graph
- Fix a threshold ϑ, delete all *vertices* (and all incident arcs) in the specified order, until ϑm arcs have been deleted
- Compute the "difference" between the graph you obtained and the original one

Experiment



Deleting nodes in order of (syntactic) depth

Experiment (cont.)



Distribution divergence (various measures)

Removal strategies compared



Removal in social networks





Findings

• Depth-order, PR etc. are strongly disruptive on web graphs

Findings

- Depth-order, PR etc. are strongly disruptive on web graphs
- Proper social networks are much more robust, still being similar to web graphs under many respects

• We propose to use spid (*shortest-paths index of dispersion*), the ratio between variance and average in the distance distribution

- We propose to use spid (*shortest-paths index of dispersion*), the ratio between variance and average in the distance distribution
- When the dispersion index is <1, the distribution is *subdispersed*; >1, is *superdispersed*

- We propose to use spid (*shortest-paths index of dispersion*), the ratio between variance and average in the distance distribution
- When the dispersion index is <1, the distribution is *subdispersed*; >1, is *superdispersed*
- Web graphs and social networks are **different** under this viewpoint!

Spid plot



spid

• We conjecture that spid is able to tell social networks from web graphs

- We conjecture that spid is able to tell social networks from web graphs
- Averag distance alone would not suffice: it is very changeable and depends on the scale

- We conjecture that spid is able to tell social networks from web graphs
- Averag distance alone would not suffice: it is very changeable and depends on the scale
- Spid, instead, seems to have a clear cutpoint at 1

- We conjecture that spid is able to tell social networks from web graphs
- Averag distance alone would not suffice: it is very changeable and depends on the scale
- Spid, instead, seems to have a clear cutpoint at 1
- What is Facebook spid?

- We conjecture that spid is able to tell social networks from web graphs
- Averag distance alone would not suffice: it is very changeable and depends on the scale
- Spid, instead, seems to have a clear cutpoint at 1
- What is Facebook spid?

[Answer: 0.093]

Average distance Effective diameter



That's all, folks!