Aalto University
School of Science

CS-E4600 – Algorithmic methods of data mining

Slide set 11 : computing basic statistics

Aristides Gionis

Aalto University

fall 2018

algorithmic tools

efficiency considerations

- data in the web and social-media are typically of extremely large scale (easily reach to billions)
- how to compute simple graph statistics?
- even quadratic algorithms are not feasible in practice

hashing and sketching

- probabilistic / approximate methods
- sketching : create sketches that summarize the data and allow to estimate simple statistics with small space
- hashing : hash objects in such a way that similar objects have larger probability of mapped to the same value than non-similar objects

graph distance distributions

small-world phenomena

small worlds : graphs with short paths

- Stanley Milgram (1933-1984)
  "The man who shocked the world"
- obedience to authority (1963)
- small-world experiment (1967)

Milgram's experiment

- 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)

Milgram's experiment

- 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)

Milgram's experiment

questions Milgram wanted to answer:

1. how many parcels will reach the target?
2. what is the distribution of the number of hops required to reach the target?
3. is this distribution different for the three starting subpopulations?

Milgram's experiment

answers to the questions

1. how many parcels will reach the target?
   29%
2. what is the distribution of the number of hops required to reach the target?
   average was 5.2
3. is this distribution different for the three starting subpopulations?
   YES: average for groups A/B/C was 4.6/5.4/5.7

chain lengths

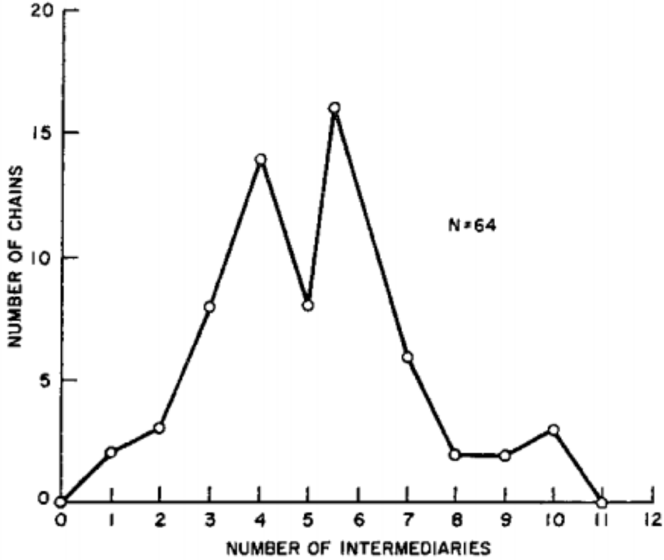


FIGURE 1

measuring what?

but what did Milgram's experiment reveal, after all?

1. the the world is small
2. that people are able to exploit this smallness

graph distance distribution

- obtain information about a large graph, i.e., social network
- macroscopic level
- distance distribution
  - mean distance
  - median distance
  - diameter
  - effective diameter
  - ...

graph distance distribution

- given a graph, $d(x, y)$ is the length of the shortest path from x to y, defined as ∞ 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 a distribution

exact computation

how can one compute the distance distribution?

exact computation

how can one compute the distance distribution?

- weighted graphs: Dijkstra (single-source: $O(m \log n)$),
- Floyd-Warshall (all-pairs: $O(n^3)$)
- in the unweighted case:
  - a single BFS solves the single-source version of the problem: $O(m)$
  - if we repeat it from every source: $O(nm)$

idea : diffusion

[Palmer et al., 2002]

- let $B_t(x)$ be the ball of radius t around x
  (the set of nodes at distance $\leq t$ from x)
- clearly $B_0(x) = \{x\}$
- moreover $B_{t+1}(x) = \bigcup_{(x,y)} B_t(y) \bigcup \{x\}$
- so computing B_{t+1} from B_t just takes a single (sequential) scan of the graph

easy but costly

- every set requires $O(n)$ bits, hence $O(n^2)$ bits overall
- easy but costly
- too many!
- what about using approximated sets?
- we need probabilistic counters, with just two primitives: add and size
- very small!

estimating the number of distinct values (F_0)

[Flajolet and Martin, 1985]

- consider a bit vector **b** with $O(\log n)$ bits
- initialize **b** to $[0, \ldots, 0]$
- consider a hash function f that maps each item x to the j-th bit of the bit-vector **b** with probability $1/2^j$
- for each item x_i in the data stream

  set the bit $j = f(x_i)$ of **b** equal to 1

  (important: bits are set deterministically for each x_i)
- let R be the index of the largest bit set
- return $Y = 2^R$

ANF

- probabilistic counter for approximating the number of distinct values [Flajolet and Martin, 1985]
- ANF algorithm [Palmer et al., 2002]
  uses the original probabilist counters
- HyperANF algorithm [Boldi et al., 2011]
  uses HyperLogLog counters [Flajolet et al., 2007]

HyperANF

- HyperLogLog counter [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

performance

- HADI, a Hadoop-conscious implementation of ANF [Kang et al., 2011]
- takes 30 minutes on a 200K-node graph
  (on one of the 50 world largest supercomputers)
- HyperANF does the same in 2.25min on a workstation
  (20 min on a laptop).

experiments on facebook

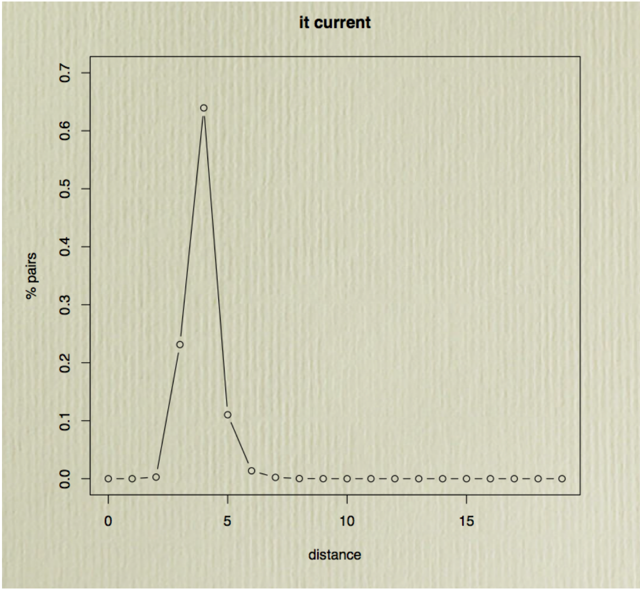
[Backstrom et al., 2011]

considered only active users

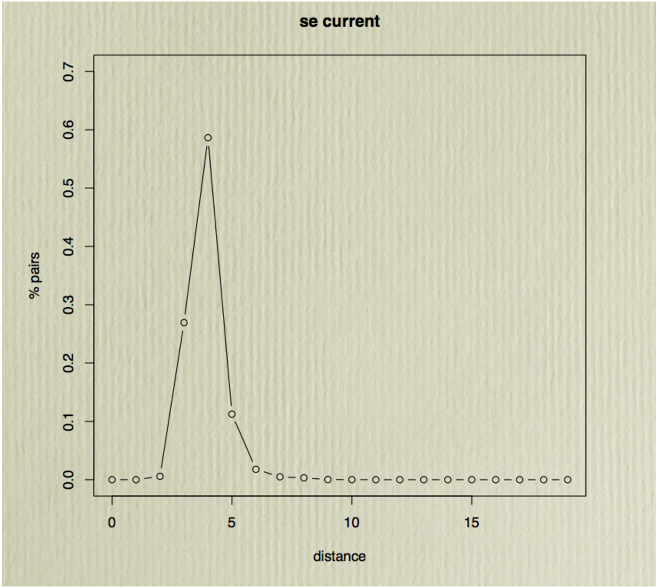
- it : only italian users
- se : only swedish users
- it + se : only italian and swedish users
- us : only US users
- the whole facebook (750m nodes)

based on users current geo-IP location

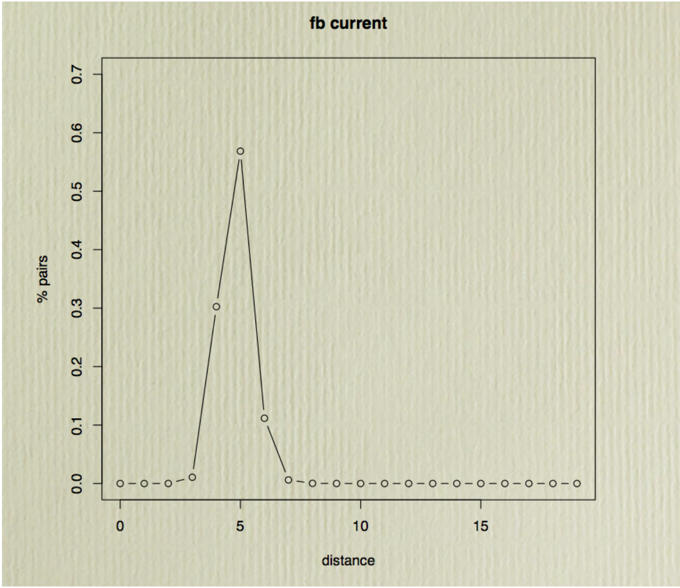
distance distribution (it)



distance distribution (se)



distance distribution (fb)



average distance

	2008	2012
it	6.58	3.90
se	4.33	3.89
it+se	4.90	4.16
us	4.74	4.32
fb	5.28	4.74

fb 2012 : 92% pairs are reachable!

effective diameter

	2008	2012
it	9.0	5.2
se	5.9	5.3
it+se	6.8	5.8
us	6.5	5.8
fb	7.0	6.2

actual diameter

	2008	2012
it	> 29	$= 25$
se	> 16	$= 25$
it+se	> 21	$= 27$
us	> 17	$= 30$
fb	> 17	> 58

breaking the news

The New York Times
Business Day
Technology

Separating You and Me? 4.74 Degrees

By JOHN MARKOFF and SOMINI SENGUPTA
Published: November 21, 2011

The world is even smaller than you thought.

Enlarge This Image

Cornell News Service
Jon Kleinberg of Cornell said weak ties could be important.

Adding a new chapter to the research that cemented the phrase "six degrees of separation" into the language, scientists at Facebook and the University of Milan reported on Monday that the average number of acquaintances separating any two people in the world was not six but 4.74.

The original "six degrees" finding, published in 1967 by the psychologist Stanley Milgram, was drawn from 296 volunteers who were asked to send a message by postcard, through friends and then friends of friends, to a specific person in a Boston suburb.

acknowledgements

Paolo Boldi Charalampos Tsourakakis

references

Backstrom, L., Boldi, P., Rosa, M., Ugander, J., and Vigna, S. (2011).
Four degrees of separation.
CoRR, abs/1111.4570.

Boldi, P., Rosa, M., and Vigna, S. (2011).
HyperANF: approximating the neighborhood function of very large graphs on a budget.
In *WWW*.

Flajolet, F., Fusy, E., Gandouet, O., and Meunier, F. (2007).
Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm.
In *Proceedings of the 13th conference on analysis of algorithm (AofA)*.

Flajolet, P. and Martin, N. G. (1985).
Probabilistic counting algorithms for data base applications.
Journal of Computer and System Sciences, 31(2):182–209.

references (cont.)

Kang, U., Tsourakakis, C. E., Appel, A. P., Faloutsos, C., and Leskovec, J. (2011).
HADI: Mining radii of large graphs.
ACM TKDD, 5.

Palmer, C. R., Gibbons, P. B., and Faloutsos, C. (2002).
ANF: a fast and scalable tool for data mining in massive graphs.
In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 81–90, New York, NY, USA. ACM Press.