Four Degrees of Separation

Lars Backstrom
Facebook
lars@fb.com

Paolo Boldi
Univ. degli Studi di Milano
boldi@dsi.unimi.it

Marco Rosa
Univ. degli Studi di Milano
marco.rosa@unimi.it

Johan Ugander
Facebook
jugander@fb.com

Sebastiano Vigna*
Univ. degli Studi di Milano
vigna@acm.org

ABSTRACT

Frigyes Karinthy, in his 1929 short story "Láncszemek" (in English, "Chains") suggested that any two persons are distanced by at most six friendship links.[1] Stanley Milgram in his famous experiments challenged people to route postcards to a fixed recipient by passing them only through direct acquaintances. Milgram found that the average number of intermediaries on the path of the postcards lay between 4.4 and 5.7, depending on the sample of people chosen. We report the results of the first world-scale social-network graph-distance computations, using the entire Facebook network of active users ($\approx$ 721 million users, $\approx$ 69 billion friendship links). The average distance we observe is 4.74, corresponding to 3.74 intermediaries or "degrees of separation", prompting the title of this paper. More generally, we study the distance distribution of Facebook and of some interesting geographic subgraphs, looking also at their evolution over time. The networks we are able to explore are almost two orders of magnitude larger than those analysed in the previous literature. We report detailed statistical metadata showing that our measurements (which rely on probabilistic algorithms) are very accurate.

INTRODUCTION

[1]The exact wording of the story is slightly ambiguous: "He bet us that, using no more than five individuals, one of whom is a personal acquaintance, he could contact the selected individual [...]". It is not completely clear whether the selected individual is part of the five, so this could actually allude to distance five or six in the language of graph theory, but the "six degrees of separation" phrase stuck after John Guare's 1990 eponymous play. Following Milgram's definition and Guare's interpretation (see further on), we will assume that "degrees of separation" is the same as "distance minus one", where "distance" is the usual path length (the number of arcs in the path).

At the 20th World–Wide Web Conference, in Hyderabad, India, one of the authors (Sebastiano) presented a new tool for studying the distance distribution of very large graphs: HyperANF [3]. Building on previous graph compression work [4] and on the idea of diffusive computation pioneered in [19], the new tool made it possible to accurately study the distance distribution of graphs orders of magnitude larger than what was previously possible.

One of the goals in studying the distance distribution is the identification of interesting statistical parameters that can be used to tell proper social networks from other complex networks, such as web graphs. More generally, the distance distribution is one interesting *global* feature that makes it possible to reject probabilistic models even when they match local features such as the in-degree distribution. In particular, earlier work [3] had shown that the *spid*[2], which measures the *dispersion* of the distance distribution, appeared to be smaller than 1 (underdispersion) for social networks, but larger than one (overdispersion) for web graphs. Hence, during the talk, one of the main open questions was "What is the spid of Facebook?".

Lars Backstrom happened to listen to the talk, and suggested a collaboration studying the Facebook graph. This was of course an extremely intriguing possibility: beside testing the "spid hypothesis", computing the distance distribution of the Facebook graph would have been the largest Milgram-like [18] experiment ever performed, orders of magnitudes larger than previous attempts (during our experiments Facebook has $\approx$ 721 million active users and $\approx$ 69 billion friendship links).

This paper reports our findings in studying the distance distribution of the largest electronic social network ever created. The average distance of the current Facebook graph is 4.74. Moreover, the spid of the graph is just 0.09, corroborating the conjecture [3] that proper social networks have a spid well below one. Contrary to what has been commonly observed analysing graphs orders of magnitude smaller, we also observe both a stabilisation of the average distance over time and that the density of the graph over time does not neatly fit previous models. Towards a deeper understanding of the structure of the Facebook graph, we apply recent compression techniques that exploit the underlying cluster

[2]The spid (shortest-paths index of dispersion) is the variance-to-mean ratio of the distance distribution.

structure of the graph to increase *locality*. The results obtained suggests the existence of overlapping clusters similar to those observed in other social networks.

Replicability of scientific results is important. While we can not release to the public the actual 30 graphs that have been studied in this paper, we distribute freely the derived data upon which the tables and figures of this papers have been built, that is, the WebGraph *properties*, which contain structural information about the graphs, and the probabilistic estimations of their neighbourhood functions (see below) that have been used to study their distance distributions. The software used in this paper is distributed under the (L)GPL General Public License.[3]

RELATED WORK

The most obvious precursor of our work is Milgram's celebrated "small world" experiment, described first in [18] and later with more details in [21]: Milgram's works were actually following a stream of research started in sociology and psychology in the late 50s [11]. In his experiment, Milgram aimed to answer the following question (in his words): "given two individuals selected randomly from the population, what is the probability that the minimum number of intermediaries required to link them is 0, 1, 2, ..., k?". In other word, Milgram is interested in computing the *distance distribution* of the acquaintance graph.

The technique Milgram used (inspired by [20]) was the following: he selected 296 volunteers (the *starting population*) and asked them to dispatch a message to a specific individual (the *target person*), a stockholder living in Sharon, MA, a suburb of Boston, and working in Boston. The message could not be sent directly to the target person (unless the sender knew him personally), but could only be mailed to a personal acquaintance who is more likely than the sender to know the target person. The starting population consisted of 100 people living in Boston chosen at random, 100 stockholders living in Nebraska (i.e., people living far from the target but sharing with him their profession), and 96 people living in Nebraska chosen at random.

In a nutshell, the results obtained from Milgram's experiments were the following: only 64 chains (22%) were completed (i.e., they reached the target); the average number of intermediaries in these chains was 5.2, with a marked difference between the Boston group (4.4) and the rest of the starting population, whereas the difference between the two other subpopulations was not statistically significant; at the other end of the spectrum, the random Nebraskan population needed 5.7 intermediaries on average (i.e., rounding up, "six degrees of separation"). The main conclusions outlined in Milgram's paper were that the average path length is small, much smaller than expected, and that geographic location seems to have an impact on the average length whereas other information (e.g., profession) does not.

Note that Milgram was measuring the average length of a

routing path on a social network, which is truly only an *upper bound* on the average distance (as the people involved in the experiment were not necessarily sending the postcard to an acquaintance on a shortest path to the destination).[4] In a sense, the results he obtained are even more striking, because not only do they prove that the world is small, but that the actors living in the small world are able to exploit its smallness. Nevertheless, it is clear that in [18, 21] the purpose of the authors is to estimate the number of intermediaries: the postcards are just a tool, and the details of the paths they follow are studied only as an artifact of the measurement process. Efficient routing was an unintended finding of these experiments, and largely went unremarked until much later [12]. Had Milgram had an actual database of friendship links and algorithms like the ones we use, we presume he would have dispensed with the postcards altogether. In the words of Milgram and Travers:

> The theoretical machinery needed to deal with social networks is still in its infancy. The empirical technique of this research has two major contribution to make to the development of that theory. First it sets an upper bound on the minimum number of intermediaries required to link widely separated Americans. Since subjects cannot always foresee the most efficient path to a target, our trace procedure must inevitably produce chains longer than those generated by an accurate theoretical model which takes full account of all paths emanating from an individual.

Thus, we believe the experiments reported in this paper are faithful to Milgram's original purpose[5], and able to overcome the problem that Milgram and Travers refer to in the above quotation—we *are* able to foresee the most efficient (shortest) path.

One difference between our experiment and Milgram's is that the notion of friendship in Facebook is hardly comparable to the idea of friendship in life; in particular, we cannot expect that all Facebook contacts are *first-name acquaintances* (as it was originally required by Milgram and Travers). This fact may artificially reduce path lengths, but also the contrary is true: since there will be many first-name acquaintances that are not on Facebook (and hence not Facebook friends) some short paths will be missing. These two phenomena will likely, at least in part, balance each other; so, although we do not have (and cannot obtain) a precise proof of this fact, we do not think we are losing or gaining much in considering the notion of Facebook friend as a surrogate of first-name friendship. Nonetheless, a strict

[4]Incidentally, this observation is at the basis of one of the most intense monologues in Guare's play: Ouisa, unable to locate Paul, the con man who convinced them he is the son of Sidney Poitier, says "I read somewhere that everybody on this planet is separated by only six other people. Six degrees of separation. Between us and everybody else on this planet. [...] But to find the right six people." Note that this fragment of the monologue clearly shows that Guare's interpretation of the "six degree of separation" idea is equivalent to distance *seven* in the graph-theoretical sense.

[5]We felt the need to state this fact very clearly, as there is a lot of confusion about this issue: in any case, we invite the reader to consult [21] directly.

comparison of Milgram's experiments and ours is difficult.[6]

We limited ourselves to the part of Milgram's experiment that is more clearly defined, that is, the estimation of the distance distribution. The largest experiment similar to the ones presented here that we are aware of is [14], where the authors considered a *communication graph* with 180 million nodes and 1.3 billion edges extracted from a snapshot of the Microsoft Messenger network; they find an average distance of 6.6 (i.e., 5.6 intermediaries; again, rounding up, six degrees of separation). Note, however, that the communication graph in [14] has an edge between two persons only if they communicated during a specific one-month observation period, and thus does not take into account friendship links through which no communication was detected.

In another study, the authors of [22] explore the distance distribution of some small-sized social networks. In both cases the networks were undirected and small enough (by at least two orders of magnitude) to be accessed efficiently in a random fashion, so the authors used *sampling* techniques. We remark, however, that sampling is not easily applicable to directed networks (such as Twitter) that are not strongly connected, whereas our techniques would still work (for some details about the applicability of sampling, see [7]).

Analyzing the evolution of social networks in time is also a lively trend of research. Leskovec, Kleinberg and Faloutsos observe in [15] that the average degree of complex networks increase over time while the *effective diameter* shrinks. Their experiments are conducted on a much smaller scale (their largest graph has 4 millions of nodes and 16 millions of arcs), but it is interesting that the phenomena observed seems quite consistent. Probably the most controversial point is the hypothesis that the number of edges $m(t)$ at time t is related to the number of nodes $n(t)$ by the following relation:

$$m(t) \propto n(t)^a,$$

where a is a fixed exponent usually lying in the interval $(1 . . 2)$. We will discuss this hypothesis in light of our findings.

DEFINITIONS AND TOOLS
The *neighbourhood function* $N_G(t)$ of a graph G returns for each $t \in \mathbf{N}$ the number of pairs of nodes $\langle x, y \rangle$ such that y is reachable from x in at most t steps. It provides data about how fast the "average ball" around each node expands. From the neighbourhood function it is possible to derive the distance distribution (between reachable pairs), which gives for each t the fraction of reachable pairs at distance exactly t.

In this paper we use HyperANF, a diffusion-based algorithm (building on ANF [19]) that is able to approximate quickly the neighbourhood function of very large graphs; HyperANF

is a probabilistic algorithm giving as output a vector random variable whose mean is the neighbourhood function to be estimated. Our implementation also uses WebGraph [4] to represent the graphs in a highly compressed but quickly accessible form.

HyperANF is based on the observation (made in [19]) that $B(x, r)$, the ball of radius r around node x, satisfies

$$B(x, r) = \bigcup_{x \to y} B(y, r - 1) \cup \{ x \}.$$

Since $B(x, 0) = \{ x \}$, we can compute each $B(x, r)$ incrementally using sequential scans of the graph (i.e., scans in which we go in turn through the successor list of each node). The obvious problem is that during the scan we need to access randomly the sets $B(x, r - 1)$ (the sets $B(x, r)$ can be just saved on disk on a *update file* and reloaded later).

The space needed for such sets would be too large to be kept in main memory. However, HyperANF represents these sets in an *approximate* way, using *HyperLogLog counters* [9], which could be thought as dictionaries that can answer just questions about size: the answer is probabilistic and depends on a random seed that is chosen independently for each run. Each counter is made of a number of small (in our case, 5-bit) *registers*, and the precision of the answer depends on the number of registers.

Theoretical error bounds
The result of a run of HyperANF at the t-th iteration is an estimation of the neighbourhood function in t. We can see it as a random variable

$$\hat{N}_G(t) = \sum_{0 \leq i < n} X_{i,t}$$

where each $X_{i,t}$ is the HyperLogLog counter that counts nodes reached by node i in t steps (n is the number of nodes of the graph). When m registers per counter are used, each $X_{i,t}$ has a guaranteed relative standard deviation $\eta_m \leq 1.06/\sqrt{m}$.

It is shown in [3] that the output $\hat{N}_G(t)$ of HyperANF at the t-th iteration is an asymptotically almost unbiased estimator of $N_G(t)$, that is

$$\frac{E[\hat{N}_G(t)]}{N_G(t)} = 1 + \delta_1(n) + o(1) \text{ for } n \to \infty,$$

where δ_1 is the same as in [9][Theorem 1] (and $|\delta_1(x)| < 5 \cdot 10^{-5}$ as soon as $m \geq 16$). Moreover, $\hat{N}_G(t)$ has a relative standard deviation not greater than that of the X_i's, that is

$$\frac{\sqrt{\text{Var}[\hat{N}_G(t)]}}{N_G(t)} \leq \eta_m.$$

In particular, our runs used $m = 64$ ($\eta_m = 0.1325$) for all graphs except for the two largest Facebook graphs, where we used $m = 32$ ($\eta_m = 0.187$). Runs were repeated so to obtain a uniform relative standard deviation for all graphs.

---

[6]Incidentally, there have been several attempts to reproduce Milgram-like routing experiments on various large networks [17, 13, 10], an independent line of research that is interesting in its own right, but we do not explore the routing properties of Facebook in this study.

EXPERIMENTS

The graphs analysed in this paper are graphs of Facebook users who were active in May of 2011; an active user is one who has logged in within the last 28 days. The decision to restrict our study to active users allows us to eliminate accounts that have been abandoned in early stages of creation, and focus on accounts that plausibly represent actual individuals. In accordance with Facebook's data retention policies, historical user activity records are not retained, and historical graphs for each year were constructed by considering currently active users that were registered by January 1st of that year, along with those friendship edges that were formed prior that that date. The "current" graph is simply the graph of active users at the time when the experiments were performed (May 2011). The graph predates the existence of Facebook "subscriptions", a directed relationship feature introduced in August 2011, and also does not include "pages" (such as celebrities) that people may "like". For standard user accounts on Facebook there is a limit of 5 000 possible friends.

We decided to extend our experiments in two directions: regional and temporal. We thus analyse the entire Facebook graph (`fb`), the USA subgraph (`us`), the Italian subgraph (`it`) and the Swedish (`se`) subgraph. We also analysed a combination of the Italian and Swedish graph (`itse`) to check whether combining two regional but distant networks could significantly change the average distance, in the same spirit as in the original Milgram's experiment.[7] For each graph we compute the distance distribution from 2007 up to today by performing several HyperANF runs, obtaining an estimate of values of neighbourhood function with relative standard deviation at most 5.8%: in several cases, however, we performed more runs, obtaining a higher precision. We report the jackknife [8] estimate of derived values (such as average distances) and the associated estimation of the standard error.

Setup

The computations were performed on a 24-core machine with 72 GiB of memory and 1 TiB of disk space.[8] The first task was to import the Facebook graph(s) into a compressed form for WebGraph [4], so that the multiple scans required by HyperANF's diffusive process could be carried out relatively quickly. This part required some massaging of Facebook's internal IDs into a contiguous numbering: the resulting current `fb` graph (the largest we analysed) was compressed to 345 GB at 20 bits per arc, which is 86% of the information-theoretical lower bound ($\log \binom{n^2}{m}$ bits for n nodes and m arcs).[9] Regardless of coding, for half of all possible

graphs with n nodes and m arcs we need at least $\lfloor \log \binom{n^2}{m} \rfloor$ bits per graph: the purpose of compression is precisely to choose the coding so to represent interesting graphs in a smaller space than that required by the bound.

To understand what is happening, we recall that WebGraph uses the BV compression scheme [4], which applies three intertwined techniques to the successor list of a node:

- successors are (partially) *copied* from previous nodes within a small window, if successors lists are similar enough;

- successors are *intervalised*, that is, represented by a left extreme and a length, if significant contiguous successor sequences appear;

- successors are *gap-compressed* if they pass the previous phases: instead of storing the actual successor list, we store the differences of consecutive successors (in increasing order) using instantaneous codes.

Thus, a graph compresses well when it exhibits *similarity* (nodes with near indices have similar successor lists) and *locality* (successor lists have small gaps).

The better-than-random result above (usually, randomly permuted graphs compressed with WebGraph occupy $10-20\%$ more space than the lower bound) has most likely been induced by the renumbering process, as in the original stream of arcs all arcs going out from a node appeared consecutively; as a consequence, the renumbering process assigned consecutive labels to all yet-unseen successors (e.g., in the initial stages successors were labelled contiguously), inducing some locality.

It is also possible that the "natural" order for Facebook (essentially, join order) gives rise to some improvement over the information-theoretical lower bound because users often join the network at around the same time as several of their friends, which causes a certain amount of locality and similarity, as circle of friends have several friends in common.

Because our computation time is greatly reduced by compression, we were interested in the first place to establish whether more locality could be induced in a graph of this size by suitably permuting the graph using *layered labelled propagation* [2] (LLP). This approach (which computes several clusterings with different levels of granularity and combines them to sort the nodes of a graph so to increase its locality and similarity) has recently led to the best compression ratios for social networks when combined with the BV compression scheme. An increase in compression means that we were able to partly understand the cluster structure of the graph.

Each of the clusterings required by LLP is in itself a *tour de force*, as the graphs we analyse are almost two orders of magnitude larger than any network used for experiments in the literature on graph clustering. Indeed, applying LLP to

---

[7] To establish geographic location, we use the users' *current* geo-IP location; this means, for example, that the users in the it-2007 graph are users who are today in Italy and were on Facebook on January 1, 2007 (most probably, American college students then living in Italy).

[8] We remark that the commercial value of such hardware is of the order of a few thousand dollars.

[9] Note that we measure compression with respect to the lower bound on *arcs*, as WebGraph stores *directed* graphs; however, with the additional knowledge that the graph is undirected, the lower

bound should be applied to *edges*, thus doubling, in practice, the number of bits used.
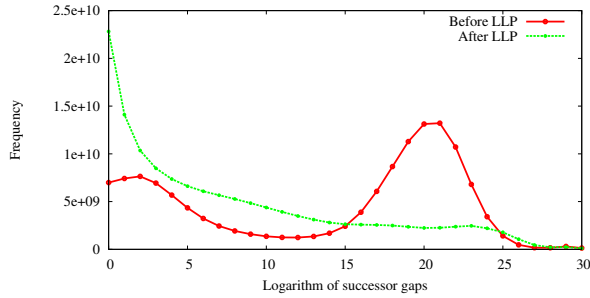
Figure 1. The change in distribution of the logarithm of the gaps between successors when the current `fb` graph is permuted by layered label propagation. See also Table 1.
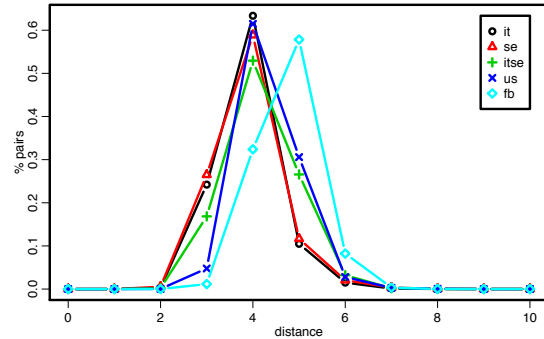

Figure 2. The probability mass functions of the distance distributions of the current graphs (truncated at distance 10).

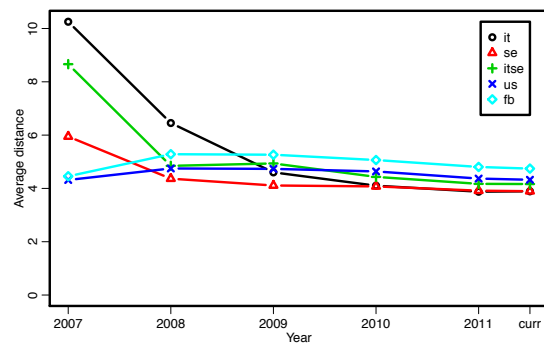**Figure 3.** The average distance graph. See also Table 6.

the current Facebook graph required ten days of computation on our hardware.

We applied layered labelled propagation and re-compressed our graphs (the current version), obtaining a significant improvement. In Table 1 we show the results: we were able to reduce the graph size by 30% (i.e., from 345 GB to 211 GB in the case of the whole Facebook graph), which suggests that LLP has been able to discover several significant clusters.

The change in structure can be easily seen from Figure 1, where we show the distribution of the binary logarithm of gaps between successors for the current `fb` graph. The smaller the gaps, the higher the locality. In the graph with renumbered Facebook IDs, the distribution is bimodal: there is a local maximum at two, showing that there is some locality, but the bulk of the probability mass is around 20–21, which is slightly less than the information-theoretical lower bound (≈ 23).

In the graph permuted with LLP, however, the distribution radically changes: it is now very nearly monotonically decreasing, with a very small bump at 23, which testifies the existence of a small core of "randomness" in the graph that LLP was not able to tame.

Regarding similarity, we see an analogous phenomenon: the number of successors represented by copying has doubled, going from 9% to 18%. The last datum is in line with other social networks (web graphs, on the contrary, are extremely redundant and more than 80% of the successors are usually copied). Moreover, disabling copying altogether results in only a modest increase in size ($\approx 5\%$), again in line with other social networks, which suggests that for most applications it is better to disable copying at all to obtain faster random access.

The compression ratio for the current `fb` graph is around 53%, which is similar to other similar social networks, such as LiveJournal (55%) or DBLP (40%) [2][10]. For other graphs

---

[10] The interested reader will find similar data for several type of networks at the LAW web site (**http://law.dsi.unimi.it/**).

(see Table 1), however, it is slightly worse. This might be due to several phenomena: First, our LLP runs were executed with only half the number or clusters, and for each cluster we restricted the number of iterations to just four, to make the whole execution of LLP feasible. Thus, our runs are capable of finding considerably less structure than the runs we had previously performed for other networks. Second, the number of nodes is much larger: there is some cost in writing down gaps (e.g., using γ, δ or ζ codes) that is dependent on their absolute magnitude, and the lower bound does not take into account that cost.

Running

The runs of HyperANF on the current whole Facebook graph used 32 registers, so the space for counters was about 27 GiB (e.g., we could have analysed a graph with four times the number of nodes on the same hardware, but in that case we would have needed a larger number of runs to obtain the same precision). As a rough measure of speed, a single run on the LLP-compressed current whole Facebook graph requires about 13.5 hours. Note that these timings would scale linearly with an increase in the number of cores.

General comments

In September 2006, Facebook was opened to non-college students: there was an instant surge in subscriptions, as our

	it	se	itse	us	fb
Original	14.8 (83%)	14.0 (86%)	15.0 (82%)	17.2 (82%)	20.1 (86%)
LLP	10.3 (58%)	10.2 (63%)	10.3 (56%)	11.6 (56%)	12.3 (53%)

Table 1. The number of bits per link and the compression ratio (with respect to the information-theoretical lower bound) for the current graphs in the original order and for the same graphs permuted by layered label propagation.
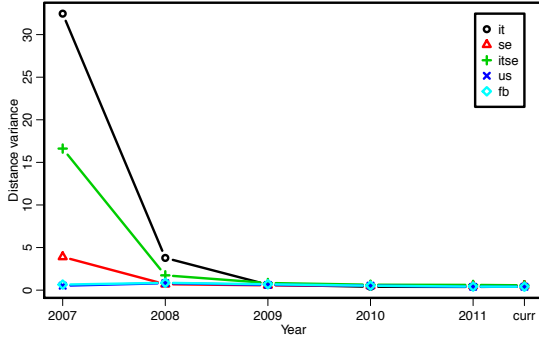

Figure 4. The graph of variances of the distance distributions. See also Table 7.

data shows. In particular, the it and se subgraphs from January 1, 2007 were highly disconnected, as shown by the incredibly low percentage of reachable pairs we estimate in Table 3. Even Facebook itself was rather disconnected, but all the data we compute stabilizes (with small oscillations) after 2009, with essentially all pairs reachable. Thus, we consider the data for 2007 and 2008 useful to observe the evolution of Facebook, but we do not consider them representative of the underlying human social-link structure.

	it	se	itse	us	fb
2007	1.31	3.90	1.50	119.61	99.50
2008	5.88	46.09	36.00	106.05	76.15
2009	50.82	69.60	55.91	111.78	88.68
2010	122.92	100.85	118.54	128.95	113.00
2011	198.20	140.55	187.48	188.30	169.03
current	226.03	154.54	213.30	213.76	190.44

Table 2. Average degree of the datasets.

	it	se	itse	us	fb
2007	0.04	10.23	0.19	100.00	68.02
2008	25.54	93.90	80.21	99.26	89.04

Table 3. Percentage of reachable pairs 2007–2008.

The distribution

Figure 2 displays the probability mass functions of the current graphs. We will discuss later the variation of the average distance and spid, but qualitatively we can immediately distinguish the *regional* graphs, concentrated around distance four, and the *whole* Facebook graph, concentrated around distance five. The distributions of it and se, moreover, have significantly less probability mass concentrated on distance five than itse and us. The variance data (Table 7

Lower bounds from HyperANF runs

	it	se	itse	us	fb
2007	41	17	41	13	14
2008	28	17	24	17	16
2009	21	16	17	16	15
2010	18	19	19	19	15
2011	17	20	17	18	35
current	19	19	19	20	58

Exact diameter of the giant component

| current | 25 | 23 | 27 | 30 | 41 |

Table 4. Lower bounds for the diameter of all graphs, and exact values for the giant component ($> 99.7\%$) of current graphs computed using the iFUB algorithm.
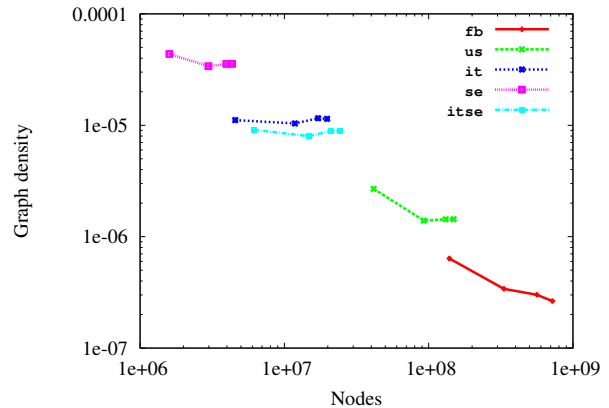

Figure 5. A plot correlating number of nodes to graph density (for the graph from 2009 on).

and Figure 4) show that the distribution quickly became extremely concentrated.

Average degree and density

Table 2 shows the relatively quick growth in time of the average degree of all graphs we consider. The more users join the network, the more existing friendship links are uncovered. In Figure 6 we show a loglog-scaled plot of the same data: with the small set of points at our disposal, it is difficult to draw reliable conclusions, but we are not always observing the power-law behaviour suggested in [15]: see, for instance, the change of the slope for the us graph.[11]

In Figure 5 we plot the *density*[12] (number of edges divided

---

[11]We remind the reader that on a log-log plot several distributions "looks like" a straight line. The quite illuminating examples shown in [16], in particular, show that goodness-of-fit tests are essential.

[12]We remark that the authors of [15] call *densification* the increase

	it	se	itse	us	fb
2007	159.8 K (105.0 K)	11.2 K (21.8 K)	172.1 K (128.8 K)	8.8 M (529.3 M)	13.0 M (644.6 M)
2008	335.8 K (987.9 K)	1.0 M (23.2 M)	1.4 M (24.3 M)	20.1 M (1.1 G)	56.0 M (2.1 G)
2009	4.6 M (116.0 M)	1.6 M (55.5 M)	6.2 M (172.1 M)	41.5 M (2.3 G)	139.1 M (6.2 G)
2010	11.8 M (726.9 M)	3.0 M (149.9 M)	14.8 M (878.4 M)	92.4 M (6.0 G)	332.3 M (18.8 G)
2011	17.1 M (1.7 G)	4.0 M (278.2 M)	21.1 M (2.0 G)	131.4 M (12.4 G)	562.4 M (47.5 G)
current	19.8 M (2.2 G)	4.3 M (335.7 M)	24.1 M (2.6 G)	149.1 M (15.9 G)	721.1 M (68.7 G)

Table 5. Number of nodes and friendship links of the datasets. Note that each friendship link, being undirected, is represented by a pair of symmetric arcs.

	it	se	itse	us	fb
2007	10.25 (±0.17)	5.95 (±0.07)	8.66 (±0.14)	4.32 (±0.02)	4.46 (±0.04)
2008	6.45 (±0.03)	4.37 (±0.03)	4.85 (±0.05)	4.75 (±0.02)	5.28 (±0.03)
2009	4.60 (±0.02)	4.11 (±0.01)	4.94 (±0.02)	4.73 (±0.02)	5.26 (±0.03)
2010	4.10 (±0.02)	4.08 (±0.02)	4.43 (±0.03)	4.64 (±0.02)	5.06 (±0.01)
2011	3.88 (±0.01)	3.91 (±0.01)	4.17 (±0.02)	4.37 (±0.01)	4.81 (±0.04)
current	3.89 (±0.02)	3.90 (±0.04)	4.16 (±0.01)	4.32 (±0.01)	4.74 (±0.02)

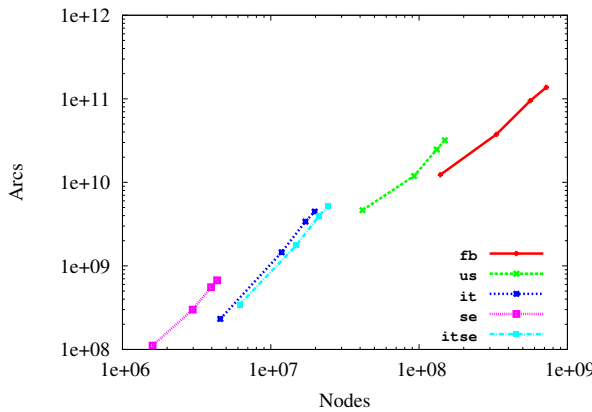**Table 6. The average distance (± standard error). See also Figure 3 and 7.**

Figure 6. A plot correlating number of nodes to the average degree (for the graphs from 2009 on).

The results concerning average distance[13] are displayed in Figure 3 and Table 6. The average distance[14] on the Facebook current graph is 4.74.[15] Moreover, a closer look at the distribution shows that 92% of the reachable pairs of individuals are at distance five or less.

On both the `it` and `se` graphs we find significantly lower but similar values. We interpret this result as telling us that the average distance is actually dependent on the geographical closeness of users, more than on the actual size of the network. This is corroborated by the higher average distance of the `itse` graph.

During the fastest growing years of Facebook our graphs show a quick decrease in the average distance, which however appears now to be stabilizing. This is not surprising, as "shrinking diameter" phenomena are always observed when a large network is "uncovered", in the sense that we look at larger and larger induced subgraphs of the underlying global human network. At the same time, as we already remarked, density was going down steadily. We thus see the small-world phenomenon fully at work: a smaller fraction of arcs

by number of nodes) of the graphs against the number of nodes. On the whole Facebook graph we see a clear trend toward sparsification; on the contrary, the density of the other regional networks seem to have stabilized, hinting possibly at the fact that the sparsification currently observed on the whole graph is due to other less mature country subnetworks and to the unavoidably lower-density connections between different countries.

In an absolute sense, though, geographical concentration increases density: in Figure 5 we can see the lines corresponding to our regional graphs clearly ordered by geographical concentration, with the `fb` graph in the lowest position.

Average distance

---

of the average degree, in contrast with established literature in graph theory, where *density* is the fraction of edges with respect to all possible edges (e.g., $2m/(n(n-1))$). We use "density", "densification" and "sparsification" in the standard sense.

[13]The data we report is about the average distance *between reachable pairs*, for which the name *average connected distance* has been proposed [5]. This is the same measure as that used by Travers and Milgram in [21]. We refrain from using the word "connected" as it somehow implies a bidirectional (or, if you prefer, undirected) connection. The notion of average distance between all pairs is useless in a graph in which not all pairs are reachable, as it is necessarily infinite, so no confusion can arise.

[14]In some previous literature (e.g., [15]), the 90% percentile (possibly with some interpolation) of the distance distribution, called *effective diameter*, has been used in place of the average distance. Having at our disposal tools that can compute easily the average distance, which is a parameterless, standard feature of the distance distribution that has been used in social sciences for decades, we prefer to stick to it. Experimentally, on web and social graphs the average distance is about two thirds of the effective diameter plus one [3].

[15]Note that both Karinthy and Guare had in mind the *maximum*, not the *average* number of degrees, so they were actually upper bounding the diameter.

	it	se	itse	us	fb
2007	32.46 (±1.49)	3.90 (±0.12)	16.62 (±0.87)	0.52 (±0.01)	0.65 (±0.02)
2008	3.78 (±0.18)	0.69 (±0.04)	1.74 (±0.15)	0.82 (±0.02)	0.86 (±0.03)
2009	0.64 (±0.04)	0.56 (±0.02)	0.84 (±0.02)	0.62 (±0.02)	0.69 (±0.05)
2010	0.40 (±0.01)	0.50 (±0.02)	0.64 (±0.03)	0.53 (±0.02)	0.52 (±0.01)
2011	0.38 (±0.03)	0.50 (±0.02)	0.61 (±0.02)	0.39 (±0.01)	0.42 (±0.03)
current	0.42 (±0.03)	0.52 (±0.04)	0.57 (±0.01)	0.40 (±0.01)	0.41 (±0.01)

Table 7. The variance of the distance distribution (± standard error). See also Figure 4.

	it	se	itse	us	fb
2007	3.17 (±0.106)	0.66 (±0.016)	1.92 (±0.078)	0.12 (±0.003)	0.15 (±0.004)
2008	0.59 (±0.026)	0.16 (±0.008)	0.36 (±0.028)	0.17 (±0.003)	0.16 (±0.005)
2009	0.14 (±0.007)	0.14 (±0.004)	0.17 (±0.004)	0.13 (±0.003)	0.13 (±0.009)
2010	0.10 (±0.003)	0.12 (±0.005)	0.14 (±0.006)	0.11 (±0.004)	0.10 (±0.002)
2011	0.10 (±0.006)	0.13 (±0.006)	0.15 (±0.004)	0.09 (±0.003)	0.09 (±0.005)
current	0.11 (±0.007)	0.13 (±0.010)	0.14 (±0.003)	0.09 (±0.003)	0.09 (±0.003)

Table 8. The index of dispersion of distances, a.k.a. spid (± standard error). See also Figure 7.

connecting the users, but nonetheless a lower average distance.

To make more concrete the "degree of separation" idea, in Table 9 we show the percentage of reachable pairs *within the ceiling of the average distance* (note, again, that it is the percentage relatively to the reachable pairs): for instance, in the current Facebook graph 92% of the pairs of reachable users are within distance five—four degrees of separation.

Spid

The *spid* is the *index of dispersion σ^2/μ* (a.k.a. *variance-to-mean ratio*) of the distance distribution. Some of the authors proposed the spid [3] as a measure of the "webbiness" of a social network. In particular, networks with a spid larger than one should be considered "web-like", whereas networks with a spid smaller than one should be considered "properly social". We recall that a distribution is called under- or over-dispersed depending on whether its index of dispersion is smaller or larger than 1 (e.g., variance smaller or larger than the average distance), so a network is considered properly social or not depending on whether its distance distribution is under- or over-dispersed.

The intuition behind the spid is that "properly social" networks strongly favour short connections, whereas in the web long connection are not uncommon. As we recalled in the introduction, the starting point of the paper was the question "What is the spid of Facebook"? The answer, confirming the data we gathered on different social networks in [3], is shown in Table 8. With the exception of the highly disconnected regional networks in 2007–2008 (see Table 3), the spid is well below one.

Interestingly, across our collection of graphs we can confirm that there is in general little correlation between the average distance and the spid: Kendall's τ is -0.0105; graphical evidence of this fact can be seen in the scatter plot shown in Figure 7. If we consider points associated with a single network, though, there appears to be some correlation between aver-
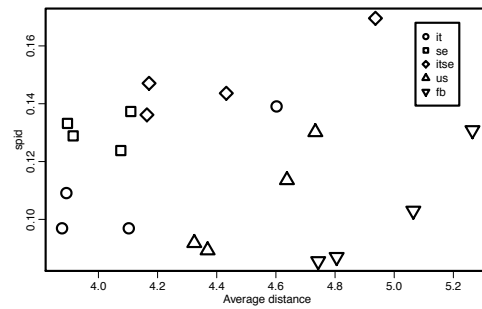

Figure 7. A scatter plot showing the (lack of) correlation between the average distance and the spid.

age distance and spid, in particular in the more connected networks (the values for Kendall's τ are all above 0.6, except for se). However, this is very likely to be an artifact, as the correlation between spid and average distance is *inverse* (larger average distance, smaller spid). What is happening is that in this case the variance (see Table 7) is changing in the same direction: smaller average distances (which would imply a larger spid) are associated with smaller variances. Figure 8 displays the mild correlation between average distance and variance in the graphs we analyse: as a network gets tighter, its distance distribution also gets more concentrated.

Diameter

HyperANF cannot provide exact results about the diameter: however, the number of steps of a run is necessarily a lower bound for the diameter of the graph (the set of registers can stabilize before a number of iterations equal to the diameter because of hash collisions, but never after). While there are no statistical guarantees on this datum, in Table 4 we report these maximal observations as lower bounds that differ significantly between regional graphs and the overall Facebook graph—there are people that are significantly more "far

	it	se	itse	us	fb
2007	65% (11)	64% (6)	67% (9)	95% (5)	91% (5)
2008	77% (7)	93% (5)	77% (5)	83% (5)	91% (6)
2009	90% (5)	96% (5)	75% (5)	86% (5)	94% (6)
2010	98% (5)	97% (5)	91% (5)	91% (5)	97% (6)
2011	90% (4)	86% (4)	95% (5)	97% (5)	89% (5)
current	88% (4)	86% (4)	97% (5)	97% (5)	91% (5)

Table 9. Percentage of reachable pairs within the ceiling of the average distance (shown between parentheses).
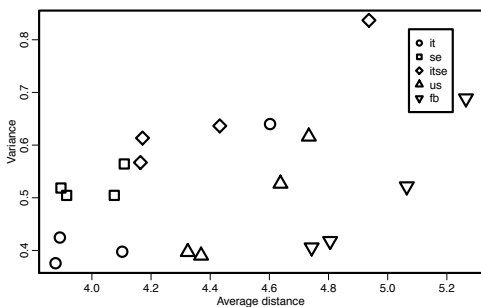

Figure 8. A scatter plot showing the mild correlation between the average distance and the variance.

apart" in the world than in a single nation.[16]

To corroborate this information, we decided to also approach the problem of computing the exact diameter directly, although it is in general a daunting task: for very large graphs matrix-based algorithms are simply not feasible in space, and the basic algorithm running n breadth-first visits is not feasible in time. We thus implemented a highly parallel version of the iFUB (iterative Fringe Upper Bound) algorithm introduced in [6].

The basic idea is as follows: consider some node x, and find (by a breadth-first visit) a node y farthest from x. Find now a node z farthest from y: $d(y, z)$ is a (usually very good) lower bound on the diameter, and actually it *is* the diameter if the graph is a tree (this is the "double sweep" algorithm).

We now consider a node c halfway between y and z: such a node is "in the middle of the graph" (actually, it would be a *center* if the graph was a tree), so if h is the eccentricy of c (the distance of the farthest node from c) we expect $2h$ to be a good upper bound for the diameter. If our upper and lower bound match, we are finished. Otherwise, we consider the *fringe*: the nodes at distance exactly h from c. Clearly, if M is the maximum of the eccentricities of the nodes in the fringe, $\max\{2(h-1), M\}$ is a new (and hopefully improved) upper bound, and M is a new (and hopefully improved) lower bound. We then iterate the process by examining fringes closer to the root until the bounds match.

Our implementation uses a multicore breadth-first visit: the

queue of nodes at distance d is segmented into small blocks handled by each core. At the end of a round, we have computed the queue of nodes at distance $d + 1$. Our implementation was able to discover the diameter of the current us graph (which fits into main memory, thanks to LLP compression) in about twenty minutes. The diameter of Facebook required ten hours of computation on a machine with 1TiB of RAM (actually, 256GiB would have been sufficient, always because of LLP compression).

The values reported in Table 4 confirm what we discovered using the approximate data provided by the length of HyperANF runs, and suggest that while the distribution has a low average distance and it is quite concentrated, there are nonetheless (rare) pairs of nodes that are much farther apart. We remark that in the case of the current fb graph, the diameter of the giant component is actually *smaller* than the bound provided by the HyperANF runs, which means that long paths appear in small (and likely very irregular) components.

CONCLUSIONS

In this paper we have studied the largest electronic social network ever created ($\approx$ 721 million active Facebook users and their $\approx$ 69 billion friendship links) from several viewpoints.

First of all, we have confirmed that layered labelled propagation [2] is a powerful paradigm for increasing locality of a social network by permuting its nodes. We have been able to compress the whole current Facebook graph at 12.3 bits per link—53% of the information-theoretical lower bound, similarly to other, much smaller social networks.

We then analyzed using HyperANF the complete Facebook graph and 29 other graphs obtained by restricting geographically or temporally the links involved. We have in fact carried out the largest Milgram-like experiment ever performed. The average distance of Facebook is 4.74, that is, 3.74 "degrees of separation", prompting the title of this paper. The spid of Facebook is 0.09, well below one, as expected for a social network. Geographically restricted networks have a smaller average distance, as it happened in Milgram's original experiment. Overall, these results help paint the picture of what the Facebook social graph looks like. As expected, it is a small-world graph, with short paths between many pairs of nodes. However, the high degree of compressibility and the study of geographically limited subgraphs show that geography plays a very significant role in forming the overall structure of network. Indeed, we see in this study, as well as

---

[16]Incidentally, as we already remarked, this is the measure that Karinthy and Guare actually had in mind.

other studies of Facebook [1] that, while the world is connected enough for short paths to exist between most nodes, there is a high degree of locality induced by various externalities, geography chief amongst them, all reminiscent of the model proposed in [12].

When Milgram first published his results, he in fact offered two opposing interpretations of what "six degrees of separation" actually meant. On the one hand, he observed that such a distance is considerably smaller than what one would naturally intuit. But at the same time, Milgram noted that this result could also be interpreted to mean that people are on average six "worlds apart": "When we speak of five[17] intermediaries, we are talking about an enormous psychological distance between the starting and target points, a distance which seems small only because we customarily regard 'five' as a small manageable quantity. We should think of the two points as being not five persons apart, but 'five circles of acquaintances' apart—five 'structures' apart." [18]. From this gloomier perspective, it is reassuring to see that our findings show that people are in fact only four world apart, and not six: when considering another person in the world, a friend of your friend knows a friend of their friend, on average.

ADDITIONAL AUTHORS

REFERENCES

1. Lars Backstrom and Eric Sun and Cameron Marlow Find me if you can: improving geographical prediction with social and spatial proximity. In *WWW*, 61–70, 2010.

2. Paolo Boldi, Marco Rosa, Massimo Santini, and Sebastiano Vigna. Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks. In *WWW*, 587–596, 2011.

3. Paolo Boldi, Marco Rosa, and Sebastiano Vigna. HyperANF: Approximating the neighbourhood function of very large graphs on a budget. In *WWW*, 625–634, 2011.

4. Paolo Boldi and Sebastiano Vigna. The WebGraph framework I: Compression techniques. In *WWW*, 595–601, 2004.

5. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. Graph structure in the Web: experiments and models. *Computer Networks 33*, 1–6 (2000), 309–320.

6. Crescenzi, P., Grossi, R., Habib, M., Lanzi, L., and Marino, A. On Computing the Diameter of Real-World Undirected Graphs. Submitted to the special issue of Theoretical Computer Science in honor of Giorgio Ausiello in the occasion of his 70th birthday., 2011.

7. Pierluigi Crescenzi, Roberto Grossi, Leonardo Lanzi, and Andrea Marino. A comparison of three algorithms for approximating the distance distribution in real-world graphs. In *TAPAS*, 92–103, 2011.

8. Bradley Efron and Gail Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.

9. Philippe Flajolet, Éric Fusy, Olivier Gandouet, and Frédéric Meunier. HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm. In *AofA 07*, 127–146, 2007.

10. Sharad Goel, Roby Muhamad, and Duncan Watts. Social search in "small-world" experiments. In *WWW*, 701–710, 2009.

11. Michael Gurevitch. *The social structure of acquaintanceship networks*. PhD thesis, Massachusetts Institute of Technology, Dept. of Economics, 1961.

12. Jon M. Kleinberg. The small-world phenomenon: an algorithm perspective. In *STOC*, 163–170, 2000.

13. Silvio Lattanzi, Alessandro Panconesi, and D. Sivakumar. Milgram-routing in social networks. In *WWW*, 725–734, 2011.

14. Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *WWW*, 915–924, 2008.

15. Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM TKDD*, 1(1), 2007.

16. Lun Li, David L. Alderson, John Doyle, and Walter Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Math.*, 2(4), 2005.

17. David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proc Nat Acad Sci USA*, 102(33):11623–11628, 2005.

18. Stanley Milgram. The small world problem. *Psychology Today*, 2(1):60–67, 1967.

19. Christopher R. Palmer, Phillip B. Gibbons, and Christos Faloutsos. Anf: a fast and scalable tool for data mining in massive graphs. In *KDD*, 81–90, 2002.

20. Anatol Rapoport and William J. Horvath. A study of a large sociogram. *Behavorial Science*, 6:279–291, 1961.

21. Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.

22. Qi Ye, Bin Wu, and Bai Wang. Distance distribution and average shortest path length estimation in real-world networks. In *ADMA*, 322–333, 2010.

---

[17]Five is the median of the number of intermediaries reported in the first paper by Milgram [18], from which our quotation is taken. More experiments were performed with Travers [21] with a slightly greater average.