GRAPH STRUCTURE OF FACEBOOK:

Module 4

Introduction

- The study of the social graph of the active users of the world's largest online social network, Facebook, mainly focused on computing *the number of users and their friendships, degree distribution, path length, clustering and various mixing patterns*.
- All calculations concerning this study were performed on Hadoop cluster using Hadoop/Hive data analysis framework.
- This social network is seen to display a broad range of unifying structural properties such as clustering, small-world effect, distribution of friends and community structure.

- Neighborhood function, denoted by $N_G(t)$ of a graph G returns for each $t \in N$, the number of pairs of vertices (x, y) such that x has a path of length at most x to y.
- It provides data about how fast the “average ball” around each vertex expands.
- It measures what percentile of vertex pairs are within a given distance.

- From this function, it is possible to derive the distance distribution which gives for each *t*, the fraction of reachable pairs at a distance of exactly *t*.

Hyper ANF Algorithm

- The Hyper ANF algorithm is a new tool for accurately studying the distance distribution of large graphs.
- It is a diffusion-based algorithm that is able to approximate quickly the neighborhood functions of very large graphs.
- It is based on the observation that B(x, r), the ball of radius r around the vertex x.

- The space needed for sets B(x,r) would be too large to be kept in main memory.
- HyperANF represents these sets in a way, using *HyperLogLog counters*, which are a kind of dictionaries that can answer questions related to size.
- Each such counter is made of a number of small *registers*.

- A register keeps track of the maximum number of trailing zeros of the values of a good hash function applied to the elements of a sequence of vertices: the number of distinct elements in the sequence is proportional to 2M.
- A technique called stochastic averaging is used to divide the stream into a number of sub streams, each analyzed by a different register.
- The result is then computed by aggregating suitably the estimation from each register.

- The main performance challenge to solve is, how to quickly compute the HyperLogLog counter associated to a union of balls, each represented by a HyperLogLog counter:
- HyperANF uses an algorithm based on word-level parallelism that makes the computation very fast.
- Another important feature of HyperANF is that it uses a *systolic approach* to avoid re computing balls that do not change during an iteration.
- This approach is fundamental to be able to compute the entire distance distribution.

- The results of a run of HyperANF at the t^{th} iteration is the estimation of the neighbourhood function in t.

$$\hat{N_G}(t) = \sum_{0 \le i < |V|} X_{i,t}$$

- where $X_{i,t}$ denotes the HyperLogLog counter that counts the vertices reached by vertex i in t steps.

- Drawbacks
  - HyperANF cannot provide exact results about the diameter.
  - The number of steps of a run is necessarily a lower bound for the diameter.

Iterative Fringe Upper Bound (iFUB) Algorithm,

- Consider some vertex *x*, and find by breadth-first search a vertex *y* farthest from *x*.
- Now, find a vertex *z* farthest from *y* by which *d(y, z)* gives us a very good lower bound on the diameter.
- Then, consider a vertex *c* halfway between *y* and *z*.
- This vertex *c* is in the middle of the graph, so if *h* is the eccentricity of *c, 2h* is expected to be a good upper bound for the diameter.
- If lower and upper bounds match, we are done.

- Otherwise, we consider the fringe: the vertices at distance exactly *h* from *c*.
- If M is the maximum of eccentricities of the vertices in fringe, max{2(h − 1), M} is a new upper bound, and M is a new lower bound.
- We then iterate the process by examining fringes closer to the root until the bounds match.
- The implementation uses a multicore breadth first search: the queue of vertices at distance d is segmented into small blocks handled by each core.
- At the end of a round, we have computed the queue of vertices at distance d + 1.

Spid- shortest-paths index of dispersion

- Measures the dispersion of degree distribution.
- Spid is defined as the variance-to-mean ratio of the distance distribution.
- It is sometimes referred to as the *webbiness* of a social network.
- Networks with spid greater than one should be considered web-like whereas networks with spid less than one should be considered properly social.

- The intuition behind this measure is that proper social networks strongly favor short connections, whereas in the web, long connections are not uncommon.
- The correlation between spid and average distance is inverse, i.e., larger the average distance, smaller is the spid.
- The spid of the Facebook graph is 0.09 thereby confirming that it is a proper social network.

Degree Distribution

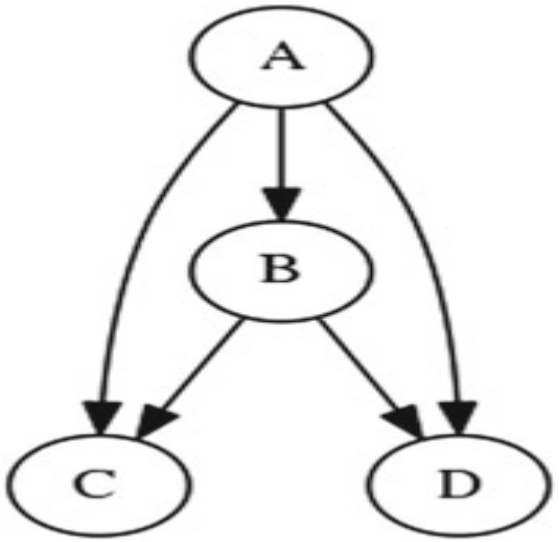
- The degree distribution of a graph G(V, E), denoted by P(k), is defined as the probability that a random chosen vertex has degree k.
- If $|V|_k$ denotes the number of vertices with degree k,

$$P(k) = \frac{|V|_k}{|V|}$$

- The degree distribution was computed by performing several HyperANF runs on the graph, obtaining estimate of the values of the neighbourhood function with relative standard deviation .
- These neighbourhood functions were computed on a single 24-core machine with 72 GiB of RAM and 1 TiB of disk space, using the HyperANF algorithm, averaging across 10 runs.

Path Length

- The length of a path is the number of edges in the sequence that comprises this path.
- The length of the path {A,B,C} in the figure is 2 because it contains the edges (A,B) and (B,C).



- The goal in studying the distance distribution is the identification of statistical parameters that can be used to tell proper social networks from other complex networks such as web graphs.
- The Facebook graph does not have paths between all pairs of vertices.
- It consists of one large connected component and therefore the neighborhood function is representative of an overwhelming majority of pair of vertices.

Component Size

- The component of a graph is a connected subgraph that is not part of any larger connected subgraph.
- The components of any graph partition its vertices into disjoint sets and the induced subgraphs of those sets.
- A K connected component ia a maximal set of vertices such that each is reachable from each of the others by at least K node independent paths.

- This component structure was analyzed using the Newman–Zipf (NZ) algorithm.
- The NZ algorithm is a type of Union-Find algorithm with path compression which records the component structure dynamically as edges are added to the network that begins completely empty of edges.
- When all the edges are added, the algorithm has computed the component structure of the network.
- This algorithm does not require that the edges must be retained in memory.
- The algorithm is applied to the Facebook graph on a single computer with 64GB of RAM by streaming over a list of edges.

Clustering Coefficient

- The clustering coefficient of a vertex *i* in a graph G(V, E), denoted by *Ci*, gives what portion of *i*'s neighbours are connected.

$$C_i = \frac{2e_i}{k_i(k_i - 1)} \epsilon [0, 1]$$

- where e_i denotes the number of edges between the neighbours of vertex *i* , k_i is the number of neighbours of a vertex
- The *average clustering coefficient* of a graph G(V, E), denoted by C, is defined as the average of the clustering coefficients of all the vertices v ∈ V.

- Neighbourhood graph for user *i* , also called the *ego* or *1-ball* graph is the graph induced subgraph consisting of users who are friends with user *i* and friendship between these users.
- The clustering coefficient decreases monotonically with degree.

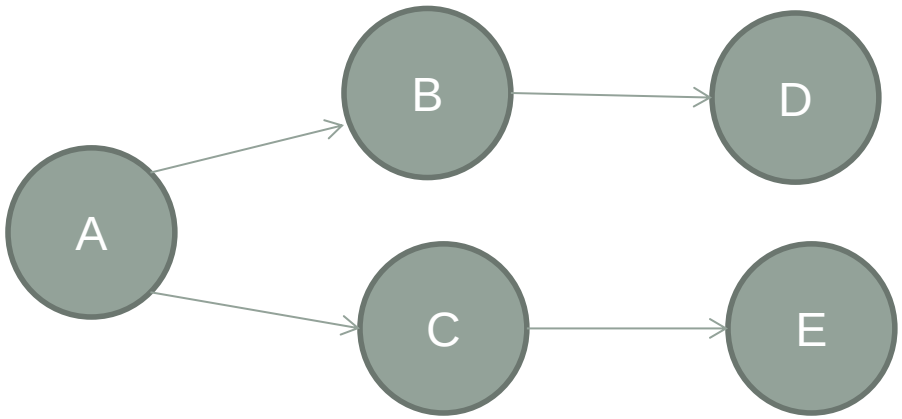
- Having observed such large clustering coefficients, the degeneracy of the graph was measured to study the sparsity of neighbourhood graphs.
- The k-core of a graph is the maximal subgraph in which all the vertices have degree at least k, i.e., the subgraph formed by iteratively removing all the vertices of degree less than k until convergence.
- The degeneracy of an undirected graph is the largest k for which the graph has a non-empty k-core.

- Even though the graph is sparse as a whole, when users accumulate sizeable friend counts, their friendships are far more indiscriminate and instead center around sizeable dense cores.

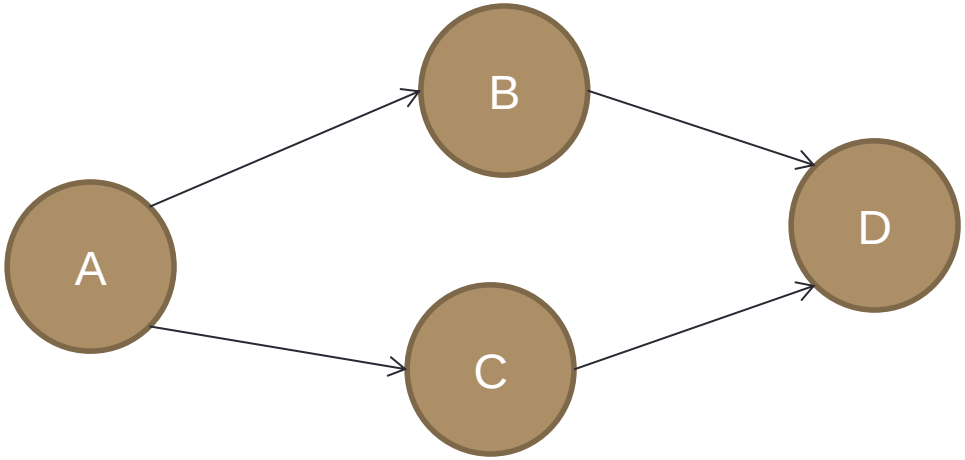
Friends-of-Friends

- The friends-of-friends denotes the number of users that are within two hops of an initial user.
- The non unique friends-of-friends count corresponds to the number of length 2 paths starting at an initial vertex and not returning to that vertex.
- The unique friends-of-friends count corresponds to the number of unique vertices that can be reached at the end of a length 2 path.

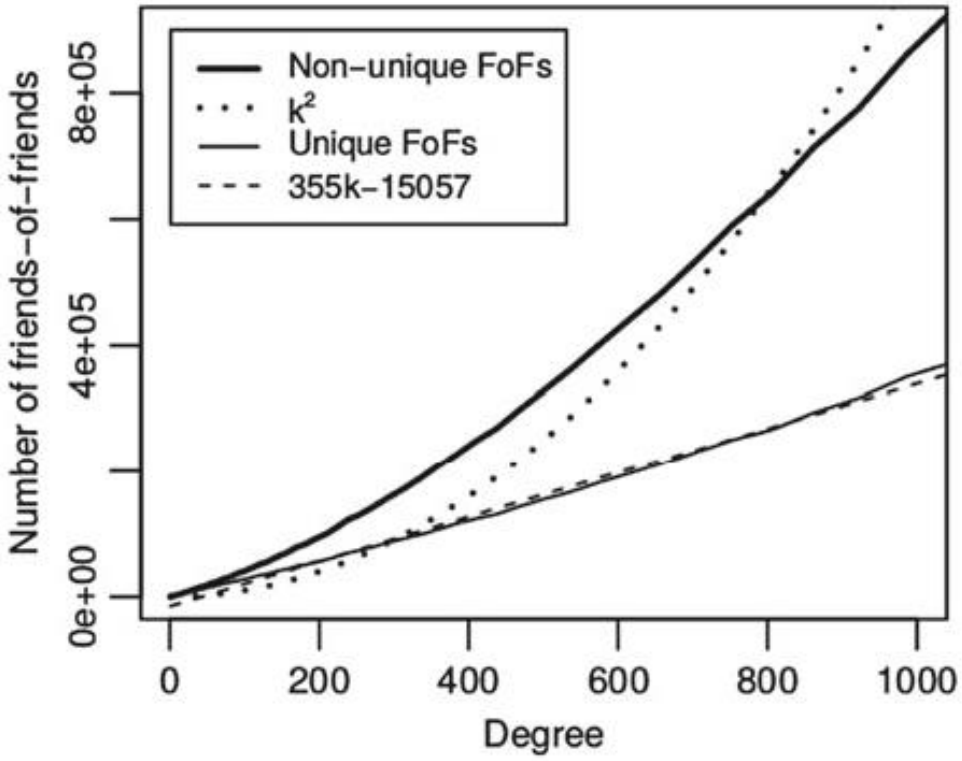
Non Unique friends-of-friends



Unique friends-of-friends



- A naive approach to counting friends-of-friends would assume that a user with k friends has roughly K^2 non-unique friends-of-friends.
- The number of unique friends-of- friends grows very close to linear, and the number of non-unique friends of- friends grows only moderately faster than linear



Average number of unique and non-unique friends-of-friends as a function of degree

Degree Assortativity

- The number of friendships in your local network neighborhood depends on the number of friends of your friends.
- Your neighbor's degree is correlated with your own degree: it tends to be large when your degree is large, and small when your degree is small.
- This is called *degree assortativity*

Login Correlation

- The definition of a random neighbour of vertices with trait x is to first select a vertex with trait x in proportion to their degree and select an edge connected to that vertex uniformly at random, i.e, we give each edge connected to vertices with trait x equal weight.
- Similar to degree’s assortativity property, login also shows a correlation between that of an individual with neighbors'.

- This correlation phenomena is best explained as follows: a Facebook user provides and receives content through status updates, links, videos, photos, etc. to and from friends, and hence may be motivated to login if they have more friends.
- So, a user who logs in more generally has more friends on Facebook and vice versa.
- So, since your friends have more friends than you do, they also login to Facebook more than you do.

Age

- To understand the friendship patterns among individuals with different ages, we compute p(t'| t) of selecting a random neighbour of individuals with age *t* who has age t'.
- A random neighbour means that each edge connected to a user with age *t* is given equal probability of being followed.
- A random neighbour is most likely to be the same age as you.
- Younger individuals have most of their friends within a small age range while older individuals have a much wider range.

Gender

- By computing p(g'| g), get the probability that a random neighbour of individuals with gender *g* has gender *g'* where *M* denotes male and *F* denotes female.
- The Facebook graph gives us the following probabilities, p(F|M) = 0.5131, p(M|M) = 0.4869, p(F|F) = 0.5178 and p(M|F) = 0.4822.
- By these computations, a random neighbour is more likely to be a female.

- There are roughly 30 million fewer active female users on Facebook with average female degree (198) larger than the average male degree (172) with p(F) = 0.5156 and p(M) = 0.4844.
- Therefore, we have $p(F|M) < p(F) < p(F|F)$ and $p(M|F) < p(M) < p(M|M)$.
- The difference between these probabilities is extremely small thereby giving a minimal effect on the preference for same gender friendships on Facebook.

Country of Origin

- An individual will have more friends from the same country of origin than from outside that country.
- The network divides along country lines into network clusters or communities.
- This division can be quantified using *modularity*, denoted by Q, *which is the fraction of edges within communities in a randomized version of the network that preserves the degree for each individual,* but is otherwise random.
- The computed value of Q = 0.7486 which is quite large and indicates a strongly modular network structure at the scale of countries.