P2P Systems and Blockchains

Spring 2018,
instructor: Laura Ricci
laura.ricci@unipi.it

Lesson 7:

WATTS STROGATZ

KLEINBERG

21/3/2018
In a random graph

- the average vertex degree depends on the probability $p$ and from the number of vertexes of the graph.

- all the nodes on the average have the same degree.

- the network diameter is low. (logarithmic)

- the clustering coefficient is low and therefore these networks are not suitable to model the aggregation, which is a characteristic of many real networks

- the low clustering coefficient enables to obtain a low value of the separation degree of the nodes.
<table>
<thead>
<tr>
<th>Network</th>
<th>Size</th>
<th>$\langle k \rangle$</th>
<th>$\ell$</th>
<th>$\ell_{\text{rand}}$</th>
<th>C</th>
<th>$C_{\text{rand}}$</th>
<th>Reference</th>
<th>Nr</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWW, site level, undir.</td>
<td>153,127</td>
<td>35.21</td>
<td>3.1</td>
<td>3.35</td>
<td>0.1078</td>
<td>0.00023</td>
<td>Adamic, 1999</td>
<td>1</td>
</tr>
<tr>
<td>Internet, domain level</td>
<td>3015–6209</td>
<td>3.52–4.11</td>
<td>3.7–3.76</td>
<td>6.36–6.18</td>
<td>0.18–0.3</td>
<td>0.001</td>
<td>Yook et al., 2001a, Pastor-Satorras et al., 2001</td>
<td>2</td>
</tr>
<tr>
<td>Movie actors</td>
<td>225,226</td>
<td>61</td>
<td>3.65</td>
<td>2.99</td>
<td>0.79</td>
<td>0.00027</td>
<td>Watts and Strogatz, 1998</td>
<td>3</td>
</tr>
<tr>
<td>LANL co-authorship</td>
<td>52,909</td>
<td>9.7</td>
<td>5.9</td>
<td>4.79</td>
<td>0.43</td>
<td>1.8×10^{-4}</td>
<td>Newman, 2001a, 2001b, 2001c</td>
<td>4</td>
</tr>
<tr>
<td>MEDLINE co-authorship</td>
<td>1,520,251</td>
<td>18.1</td>
<td>4.6</td>
<td>4.91</td>
<td>0.066</td>
<td>1.1×10^{-5}</td>
<td>Newman, 2001a, 2001b, 2001c</td>
<td>5</td>
</tr>
<tr>
<td>SPIRES co-authorship</td>
<td>56,627</td>
<td>173</td>
<td>4.0</td>
<td>2.12</td>
<td>0.726</td>
<td>0.003</td>
<td>Newman, 2001a, 2001b, 2001c</td>
<td>6</td>
</tr>
<tr>
<td>NCSTRL co-authorship</td>
<td>11,994</td>
<td>3.59</td>
<td>9.7</td>
<td>7.34</td>
<td>0.496</td>
<td>3×10^{-4}</td>
<td>Newman, 2001a, 2001b, 2001c</td>
<td>7</td>
</tr>
<tr>
<td>Math. co-authorship</td>
<td>70,975</td>
<td>3.9</td>
<td>9.5</td>
<td>8.2</td>
<td>0.59</td>
<td>5.4×10^{-5}</td>
<td>Barabási et al., 2001</td>
<td>8</td>
</tr>
<tr>
<td>Neurosci. co-authorship</td>
<td>209,293</td>
<td>11.5</td>
<td>6</td>
<td>5.01</td>
<td>0.76</td>
<td>5.5×10^{-5}</td>
<td>Barabási et al., 2001</td>
<td>9</td>
</tr>
<tr>
<td>E. coli, substrate graph</td>
<td>282</td>
<td>7.35</td>
<td>2.9</td>
<td>3.04</td>
<td>0.32</td>
<td>0.026</td>
<td>Wagner and Fell, 2000</td>
<td>10</td>
</tr>
<tr>
<td>E. coli, reaction graph</td>
<td>315</td>
<td>28.3</td>
<td>2.62</td>
<td>1.98</td>
<td>0.59</td>
<td>0.09</td>
<td>Wagner and Fell, 2000</td>
<td>11</td>
</tr>
<tr>
<td>Ythan estuary food web</td>
<td>134</td>
<td>8.7</td>
<td>2.43</td>
<td>2.26</td>
<td>0.22</td>
<td>0.06</td>
<td>Montoya and Solé, 2000</td>
<td>12</td>
</tr>
<tr>
<td>Silwood Park food web</td>
<td>154</td>
<td>4.75</td>
<td>3.40</td>
<td>3.23</td>
<td>0.15</td>
<td>0.03</td>
<td>Montoya and Solé, 2000</td>
<td>13</td>
</tr>
<tr>
<td>Words, co-occurrence</td>
<td>460,902</td>
<td>70.13</td>
<td>2.67</td>
<td>3.03</td>
<td>0.437</td>
<td>0.0001</td>
<td>Ferrer i Cancho and Solé, 2001</td>
<td>14</td>
</tr>
<tr>
<td>Words, synonyms</td>
<td>22,311</td>
<td>13.48</td>
<td>4.5</td>
<td>3.84</td>
<td>0.7</td>
<td>0.0006</td>
<td>Yook et al., 2001b</td>
<td>15</td>
</tr>
<tr>
<td>Power grid</td>
<td>4941</td>
<td>2.67</td>
<td>18.7</td>
<td>12.4</td>
<td>0.08</td>
<td>0.005</td>
<td>Watts and Strogatz, 1998</td>
<td>16</td>
</tr>
<tr>
<td>C. Elegans</td>
<td>282</td>
<td>14</td>
<td>2.65</td>
<td>2.25</td>
<td>0.28</td>
<td>0.05</td>
<td>Watts and Strogatz, 1998</td>
<td>17</td>
</tr>
</tbody>
</table>
In the figure presented in the previous slide:

- \textit{size} = network size
- \textit{<k>} = average node degree
- \textit{L} = average length of the path between two nodes
- \textit{C} = clustering coefficient
- \textit{l_{\text{rand}}} = average length of the path between pair of nodes in a random graph of the same size, where the average degree of nodes is the same
- \textit{C_{\text{rand}}} = average clustering coefficient of a random graph of the same size, whose average node degree is the same
SMALL WORLD NETWORKS

- What can you see from the previous slides?

- Many real networks are characterized by a very low diameter.

- In several social networks, individuals tend to group in clusters so that the network is characterized by a high clustering coefficient.

- Erdos-Reny graphs are characterized by a low average path length, but not by a high clustering coefficient.

- Problem: is it possible to define more realistic models for complex real networks?

- **Watts Strogatz Model**: a model suitable for describing small words networks.
REGULAR NETWORKS: PROPERTIES

- d-dimensional k-regular grids (d-lattice networks)
  - d: dimensions, k: number of neighbours
  - $d=1, k=2$: ring
  - $d=2, k=4$: mesh

- opposite characteristics with respect to a random graph
  - high diameter, but also high clustering coefficient
another network model: d-dimensional k-regular grids (d-lattice networks)

- opposite characteristics with respect to a random graph
  - high diameter, but also high clustering coefficient
How describe the structure of a social network or, peer to peer semantic overlay: the links represents connections between peers with common interests.

Intuition: each person has a set of knowledges regarding the “close acquaintances”: the colleagues, the classmates, the neighbours,..

We may think that the resulting network has “a grid structure”, and has high clustering coefficient.

Experimental results show that these network also have a small diameter.

Looking for a model generating networks with both: small diameter high clustering coefficient.
Watts-Strogatz considers two fundamental characteristics of the complex networks:

- clustering coefficient
- distance between the vertexes.

If the clustering coefficient is high, the average distance between two nodes would be high, because the edges are “local”

But.....most networks present an high value of the clustering coefficient (0.3-0.4) and a low value of the node distance

Find a scientific justification for scenarios like that describe the following conversation

- hello, I have moved to Lucca
- ah yes, then, perhaps, you know Mario Rossi?
- yes, I know. Of course the world is small....
BETWEEN ORDER AND CHAOS

- Watts-Strogatz basic idea:
  - a grid-like network presents a high aggregation, but it is not a small world network
  - random graph has a low clustering coefficient proportional to the probability of defining an edge between two nodes
  - a model with both characteristics:
    - a sufficient degree of clustering = high clustering coefficient
    - a sufficient degree of chaos to have a low separation degree among the nodes
  - a compromise between order and chaos .. a small world model
Watts-Strogatz proposes an **hybrid model**

- starts from a ring of \( n \) vertexes
- connects each vertex with its \( k \)-nearest neighbours
- “rewires” each edge with a **probability** \( p \)
  - given an edge, fix a vertex of an edge and choose another target for the other vertex, uniformly at random among all other vertexes
The resulting network model properly describes a network where

- several vertexes have links with close neighbours.
- some vertex have a link to a long range neighbour: in social networks these links corresponds to links connecting different communities.
MODELLING REAL NETWORKS

Local edges (many)

Long-range edges (few random shortcuts)

High clustering

Short paths
THE WATTS STROGATZ MODEL

N = 20, k = 8

- when probability $p=0$, the resulting network is completely regular, with a clustering coefficient of about $\frac{3}{4}$ for large values of $k$ and a diameter $O(n)$
- when $p=1$, the resulting network may be considered a random graph with a low clustering coefficient basso and a diameter $O(\log n)$
- interesting scenarios: intermediate values of the probability $p$
**THE WATTS-STROGATZ MODEL**

- **Watts-Strogatz procedure:** Let $V=\{v_1,v_2,\ldots,v_n\}$ be the vertexes of a graph. Let $k$ be an even value. Let $n >> k >> \ln(n) >> 1$
  - order a vertex on a ring
  - connect each vertex to its first $k/2$ left neighbours on the ring (clockwise) and to its first $k/2$ right neighbours (anti-clockwise). This defines the starting graph $G$.
  - with probability $p$, substitute an edge $\langle u,v \rangle$ with an edge $\langle u,w \rangle$, where $w \neq v$ is chosen uniformly at random among the vertexes of the graph which is modified, such that $\langle u,w \rangle \notin E(G)$.
- $n$ must be larger than $k$, otherwise the links with the neighbours covers all the vertexes in the graph
- $WS(n,k,p)$: a Watts e Strogatz graph with $n$ vertexes, $k$ neighbours, probability $p$ of edges rewiring.
**WS(n,k,p): CLUSTERING COEFFICIENT**

- the clustering coefficient for $WS(n,k,0)$ has been computed in the previous lesson (see the exercise)

- the clustering coefficient for $WS(n,k,p)$ is the following one:

  $$CC\ (WS(n,k,p)) = CC(WS(n,k,0)) \times (1-p)^3$$

- an intuition about the proof:
  - the probability that the rewiring process forms a triangle is very low
  - the triangles existing before the application of the rewiring process mostly contribute to the rewiring process, a few new triangles are created
  - let us consider a triangle in $WS(n,k,0)$, a triangle existing before the rewiring process
    - the probability that it continues to exist after the rewiring procedure is equal to the probability that no one of the 3 edges is affected by the rewiring, that is $(1-p)^3$
The clustering coefficient of a graph $WS(n,k,0)$ is high also when the value of $k$ is low, but...

Theorem: $\forall G \in WS(n,k,0)$, the average length of the shortest path may be approximated as follows:

$$d(u) \approx \frac{(n-1)(n+k-1)}{2kn}$$

$WS(n,k,0)$
- presents an high clustering coefficient
- cannot be considered a small world, because the average length of the shortest paths remains high
- but...for low values of $p$ (rewiring probability) the average path length rapidly decreases, while the clustering coefficient remains high.
The diagram shows the clustering coefficient and the average distance as a function of the probability $p$, in Watts Strogatz.

- The clustering coefficient is high for small values of $p$, but the average path length rapidly decreases.

- Conclusion: $WS(n,k,p)$ for small values of $p$ is a small world.
WATTS-STROGATZ: INFORMATION DIFFUSION
WATTS-STROGATZ: INFORMATION DIFFUSION
Random Graphs:
- The average lengths of the shortest paths is $O(\log(n))$, where $n$ is the number of nodes in the network.
- low clustering coefficient

Watts e Strogatz:
- defines a regular network
  - grid
  - ring with links to the nodes whose distance is less or equal $k$.
- “overlaps” a set of links generated uniformly at random to this regular structure

Conclusion: A few links generated at random put in a highly clustered graph are able to generate paths whose length is comparable with that of a random graph, while the underlying structure defines a good level of clustering.
THE ORACLE OF BACON: A SMALL WORLD

- film collaboration network: the links connect two actors which have played in at least one film  (internet movies database)

- A “demo” showing that this network is a small world network can be accessed at the URL http://oracleofbacon.org/, through the Kevin Bacon game.

- Kevin Bacon Game:
  - think of an actor A
  - if A has played in a film with Bacon, A has a Bacon number = 1
  - if A has never played with Bacon, but has played with someone hat has played, in turn, with Bacon,  A has Bacon number =2
  - and so on....
Examples:

- Marcello Mastroianni: Bacon Distance 2
  - Marcello Mastroianni in *Poppies Are Also Flowers* (1966) with Eli Wallach
  - Eli Wallach in *Mystic River* (2003) with Kevin Bacon
- Brad Pitt: Bacon Distance 1
  - Brad Pitt was in *Sleepers* (1996) with Kevin Bacon
- Elvis Presley: Bacon Distance 2
  - Elvis Presley in *Live a Little, Love a Little* (1968) with John Wheeler
  - John Wheeler in *Apollo 13* (1995) with Kevin Bacon
# The Oracle of Bacon

## How good a center is Kevin Bacon?

<table>
<thead>
<tr>
<th>Kevin Bacon Number</th>
<th># of People</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2,349</td>
</tr>
<tr>
<td>2</td>
<td>22,3940</td>
</tr>
<tr>
<td>3</td>
<td>666,941</td>
</tr>
<tr>
<td>4</td>
<td>153,220</td>
</tr>
<tr>
<td>5</td>
<td>9,652</td>
</tr>
<tr>
<td>6</td>
<td>877</td>
</tr>
<tr>
<td>7</td>
<td>134</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
</tr>
</tbody>
</table>

Total number of linkable actors: 105,713,900

Weighted total of linkable actors: 31,185,620

Average Kevin Bacon number: 2.950
THE ORACLE OF BACON: A SMALL WORLD

- Data about actors and films come from Internet Movie Database of the University of Virginia

- Each actor is separated from Kevin Bacon by a few links

- The average of Bacon Number is 2.78

- Kevin Bacon, an actor not so famous, seems to be at the centre of a collaborations network between actors, but....this is not completely true

- Kevin Bacon has a limited number of links with other actors, has played in a limited set of films, but the average distance of an actor from Bacon (2.78) is rather high
rodolfo valentino has a martina stella number of 3.

Rudolph Valentino

Character Studies (1927)

Jackie Coogan

John Goldfarb, Please Come Home (1955)

Carl Reiner

Ocean's Twelve (2004)

Martina Stella
## ACTORS CENTRALITY (212250 ATTORI)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Average distance</th>
<th>#of movies</th>
<th># of links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rod Steiger</td>
<td>2.537527</td>
<td>112</td>
<td>2562</td>
</tr>
<tr>
<td>2</td>
<td>Donald Pleasence</td>
<td>2.542376</td>
<td>180</td>
<td>2874</td>
</tr>
<tr>
<td>3</td>
<td>Martin Sheen</td>
<td>2.551210</td>
<td>136</td>
<td>3501</td>
</tr>
<tr>
<td>4</td>
<td>Christopher Lee</td>
<td>2.552497</td>
<td>201</td>
<td>2993</td>
</tr>
<tr>
<td>5</td>
<td>Robert Mitchum</td>
<td>2.557181</td>
<td>136</td>
<td>2905</td>
</tr>
<tr>
<td>6</td>
<td>Charlton Heston</td>
<td>2.566284</td>
<td>104</td>
<td>2552</td>
</tr>
<tr>
<td>7</td>
<td>Eddie Albert</td>
<td>2.567036</td>
<td>112</td>
<td>3333</td>
</tr>
<tr>
<td>8</td>
<td>Robert Vaughn</td>
<td>2.570193</td>
<td>126</td>
<td>2761</td>
</tr>
<tr>
<td>9</td>
<td>Donald Sutherland</td>
<td>2.577880</td>
<td>107</td>
<td>2865</td>
</tr>
<tr>
<td>10</td>
<td>John Gielgud</td>
<td>2.578980</td>
<td>122</td>
<td>2942</td>
</tr>
<tr>
<td>11</td>
<td>Anthony Quinn</td>
<td>2.579750</td>
<td>146</td>
<td>2978</td>
</tr>
<tr>
<td>12</td>
<td>James Earl Jones</td>
<td>2.584440</td>
<td>112</td>
<td>3787</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>876</td>
<td>Kevin Bacon</td>
<td>2.786981</td>
<td>46</td>
<td>1811</td>
</tr>
</tbody>
</table>
THE NAVIGATION PROBLEM

- Suppose you are the node of a large technological network
  - peer to peer network
  - a social network
  - the network of Internet routers

- the node is looking for information stored in another node or for a friend in a social network

- the node does not know where is the information it is looking for
  - it does not have the whole graph of the network
  - it has limited information on the target
  - it only knows its own neighbours
  - it want to find the information efficiently

- we refer to this problem as “distributed navigation problem”
THE NAVIGATION PROBLEM

- Two aspects for solving the navigation (search) problem:
  - **structural issue**: some short path must exist in the network, otherwise the information cannot be retrieved efficiently.
  - **algorithmic issue**: allow nodes to find these short paths using only distributed, local information.

- Several experiments in real world have shown that short paths exist and it is possible to find them by a distributed search.

- The Watts Strogatz model
  - allows to build a network short paths and high clustering
  - but is it a good model to build a distributed algorithm that finds these short paths in a WS graph, by exploiting only local information?
THE MILGRAM EXPERIMENT

An experiment made in 1969 by Travers (Harvard) and Milgram (City University, New York)

The abstract of the paper:

“Arbitrarily selected individuals (N=296) in Nebraska and Boston are asked to generate acquaintance chains to a target person in Massachussets, employing “the small world method” [Milgram 1967]. 64 chains reach the target person. Within this group the mean number of intermediaries between starters and targets is 5.2. Boston starting chains reach the target person with fewer intermediaries than those starting in Nebraska; sub-populations in the Nebraska group do not differ among themselves. The funneling of chains through sociometric "stars" is noted, with 48 per cent of the chains passing through three persons before reaching the target. Applications of the method to studies of large scale social structure are discussed”
THE MILGRAM EXPERIMENT

- arbitrarily select a group of starters in Nebraska and Boston area and a target person in Massachusetts
- give each starter a letter and ask he/she to move the letter toward the target
  - information about the name of the target (a broker in Boston) was in the letter, but the exact address of the trader is not known
- the letter could be forwarded only to first-name acquaintance sender
  - Alice must deliver a letter to Bob, Alice does not know Bob, she chooses her friend Mary, because she thinks that Mary has a greater probability to reach Bob.
- acquaintance chain ends when the target is reached or when someone declines to participate
goal of the research of Milgram:

structural:

- how “far” are two person chosen at random?
- distance is defined as knowledge chain “A knows B”, “B knows C”, “C knows D”,..
- what is the degree of separation between the two persons?

algorithmic:

- are these person able to find each other without knowing the whole social relationships?

- each person:
  - chooses among the long range links the one reducing the distance with respect to the target
  - exploits a local knowledge incorporating an intrinsic knowledge of the network

![Diagram of Small World Networks]

---

Laura Ricci
Watts Strogatz and Kleinberg
THE MILGRAM EXPERIMENT: THE RESULTS

• Only 64 of the initial 296 chains ended

• Milgram computed the average number of "transfers” of each letter which reaches the receiver

• Six degree of separation: each letter reaches the destination by employing no more than 6 passages: 5.2 steps on the average are required to reach the destination.

• the experiment showed that
  • the diameter of the social network is very low even if the network is highly clustered.
  • the persons are able to detect these paths in a completely decentralized way, only having local knowledge, without having a map of the complete network

• Navigable small world networks
LENGHT DISTRIBUTION OF COMPLETED CHAINS

![Graph showing the length distribution of completed chains. The x-axis represents the number of intermediaries, ranging from 0 to 12. The y-axis represents the number of chains, ranging from 0 to 20. The graph peaks at 6 and 7 intermediaries, with a total of 64 chains.](image-url)
Issues empirically detected from the Milgram experiment

- not only the “short paths” exist, but the person are able to discover these paths through a local (partial) view of the network
- long-range links in the first steps, then smaller hops
THE MILGRAM EXPERIMENT: THE RESULTS

- Many of the completed chains passed through a very small number of penultimate individuals: “funnels”
- An individual responsible for forwarding 16% (out of 64) chains to the target
- Other two individuals responsible for 10 and 5 chains, respectively.
- Connectors” or “hubs” with high degree often exist in social networks
- Target need not be a “connector” for small world phenomenon to exist
- Like “hub” airports in air traffic
Several further experiments in the last 10 years

- **Columbia Small World Project**: web based, e-mail tracking
  - online registration of participants, electronic tracking
  - average completed chain length = 4.05

- **Microsoft Instant Messenger study**:
  - only a structural study
  - two users connected only if they communicated during last month
  - graph is fully known
  - single giant component
  - average distance = 6.6

- **Small World Facebook study**:
  - 2012 structural study of Facebook relationships
  - average path length: 4.74
• Watts and Strogatz is navigable? is it possible to discover the shortest paths in a WS graph by a distributed algorithm able to discover “the shortest paths”?

• a simple approach: a greedy algorithm
  • a node knows its own neighbours and their “geographic location”
  • next hop is the “geographically closest” to the destination
  • this algorithm is able to find the “shortest paths” in a logarithmic time?
• WS is not navigable: no local greedy algorithm is able to find out the shortest paths, even if they exist!

• what is the problem? Consider the blue circled node:
  • it looks for the neighbour y which is closer to the destination
  • even if y is closer to the destination with respect to x, it is not possible to foresee the location of the neighbours of y
  • all the neighbours of y could be farther than x with respect to the destination node, dest
THE NAVIGABILITY PROBLEM

- In the WS, the random links are not correlated to the “geographic structure” of the network, but are uniformly distributed over the whole grid the random links are “too random”

- The long range links have to be not completely random, but distributed at random on different distance bands
  - uniform on different bands of distances

- A new model: the Kleinberg model
  - consider this issue, by changing the definition of the long range links in order to
    - “incorporate” the short paths into the network.
    - find short paths by a decentralized search procedure
THE KLEINBERG MODEL

- networks with a low diameter and high level of clustering, like Watts Strogatz
- ...but, the remote contacts are defined by considering the geometry of the network
  - this implies that the network is navigable
  - it is possible to define a decentralized routing algorithm
- exploited as a basic model to define the following P2P overlays
  - Freenet
  - Symphony
  - SWOP
  - Viceroy
THE KLEINBERG MODEL: DEFINITION

- consider a set of nodes corresponding to the points of a n-dimensional grid

- distance on the grid between the nodes \((i,j)\) e \((k,l)\)
  - Manhattan distance: \(d((i,j), (k,l)) = |k-i| + |l-j|\)

- given a constant \(m \geq 1\), for each node \(u\), define local contacts
  - edges directed from \(u\) to any node that is located within a distance \(m\).

- given a constant \(k \geq 0\), for each node \(u\), define \(k\) remote contacts, \(k\) edges directed from \(u\) to targets defined such that:
  - the probability to define a remote contact with target the node \(v\) distant \(d(u,v)\) from \(u\) is proportional to \(1/d(u,v)^\alpha\), where \(\alpha\) is the clustering exponent, which defines the model navigability

- remote contacts are generated so that, approximatively the same number of contacts is generated for each distance band
  - different degrees of navigability are obtained by modifying the value of \(\alpha\)
THE KLEINBERG MODEL

A bidimensional grid, $m=1$, $k=2$
THE KLEINBERG MODEL

**Local edges** $(p)$

$$A \rightarrow E := d(A,E) \leq p$$

**Long-range edges** $(q)$

$$\Pr(A \rightarrow Z) \sim 1/[d(A,Z)]^\alpha$$

*Inverse $\alpha^{th}$-power distribution*

**Lattice distance**

$$d(A,Z) = |t-u| + |w-v|$$
THE KLEINBERG MODEL: NAVIGABILITY

- consider a grid of dimension $N$

- $\alpha = 0$
  - links are chosen uniformly at random on all the network: WS model

- $\alpha > N$
  - a few long range links
  - most links with close neighbours,
  - the decentralized search
    - rapidly finds a target in the neighbourhood
    - if the target is farther, the search of the target is “slow”

$N = 2, \alpha = 4$
• consider a grid of dimension $N$

• $\alpha < N$
  • many long range links
  • the decentralized search
    • rapidly approaches the target neighbourhood
    • slows down till the target is reached

• green edges
  • bring to alternative paths
  • no one brings toward the right direction, to the target

\[ N = 2, \ \alpha = 0 \]
\( \alpha = 2 \)

the search progresses on distances bands which are farther and farther from the origin

at each distance band
- a set of hops through the local contacts
- a change of distance band

\( \alpha = N: \) optimal neighbourhood navigability

\( N = 2, \alpha = 2 \)

\( p \sim \frac{1}{d^2} \)
Theorem:

Let $d(v,w)$ be the distance between two nodes in a bi-dimensional grid of $n$ nodes. If two vertexes $v$ and $w$ are connected by a long range link with probability $1/d(u,v)^\alpha$, with $\alpha=2$, then it is possible to define a decentralized routing algorithm $A$ and a constant $\alpha_2$, independent from $n$, such that when $m=1$ and $k=1$, the number of steps required by $A$ to transmit a message between a pair of nodes is at most $\alpha_2 (\log n)^2$.

$\alpha$: clustering exponent characterizing the model.
• The proof of the theorem is rather complex, we can try to give an intuitive explanation.

• Given a node $n$, consider a series of “incremental distance band” with respect to $n$.

• For each distance band, approximatively define the same number of links.

• If a band $f_1$ is farther than a band $f_2$, it includes a larger set of nodes, but the probability to define a link with the nodes in that distance band decreases, so the procedure approximatively generates the same number of links for each distance band.
THE KLEINBERG MODEL: INTUITION

- Given a node \( v \) and a fixed distance \( d \), let us consider the nodes which are located within a distance band between \( d \) e \( 2d \)
  - the total number of nodes is proportional to \( d^2 \)
  - the probability that \( v \) defines a long range link toward one of these nodes is inversely proportional with respect to \( d^2 \)

- The two terms cancel each other (approximatively).
- The number of long range links directed to a node at distance between \( d \) e \( 2d \) is independent from \( d \)
  - the long range links are uniformly distributed on the different distance bands
THE KLEINBERG MODEL

• given a node $u$, partition the remaining nodes into the sets $A_1, A_2, A_3, \ldots A_n$, corresponding to different distance bands
• the larger the distance, a larger number of nodes is located in that distance band, but the probability of selecting one of them is smaller
• if $\alpha = d$, the long range links of $u$ are uniformly distributed on the different sets $A_i$. 
If the contacts are distributed “a la Kleinberg”

• it is possible to exploit a greedy routing, with the guarantee that the target is reached in a logarithmic number of steps.

• greedy routing navigation:

  “At each step, choose among local and remote contacts the contact which is closer to the target, according to the metrics exploited to measure the distance on the grid”
HOW MANY LONG RANGES?

- Let us consider a Kleinberg network of d dimension with clustering exponent $a = d$

- the search time depends also on the number of long range links ($k$):

  - with a single long range link ($k = 1$)
    - average search time is $O(\log^2 n)$
  
  - with $c$ long range links, ($k = c$)
    - average search time is $O(\log^2 n)/c$

  - with $k=\log n$ long range links,
    - the average search time is $O(\log n)$
    - same complexity bound of Chord is obtained in this case!
CHORD VERSUS KLEINBERG

- several structured P2P overlay are derived from the Kleinberg model
- a randomized version of Chord
  - each finger is chosen at random on different distance bands
THE KLEINBERG MODEL: CONCLUSIONS

- with respect to Watts Strogatz
  - Watts e Strogatz procedure
    - generates a small world network
    - does not enable the definition of an effective and efficient routing algorithm
  - Kleinberg generates a small world network which is navigable

- With respect to a classical P2P overlay
  - if \( k = \log(n) \), \( m = 1 \) Kleinberg generates a network
    - similar to Chord
    - with a random choice of the links.
  - is able to describe also non structured networks like Gnutella
SYMPHONY: A KLEINBERG BASED DHT


- the structure of Symphony is similar to that of Chord:
  - pairs an identifier in the logical space to each node and each data.
  - identifiers are assigned in the interval $[0,1]$, in a unitary ring
  - segment of responsibility: each node is responsible of all the data identified by
    - an identifier larger or equal to its identifier (clockwise ordering) and less than or equal to the identifier of the next node
  - but....long range links are put at random, according Kleinberg's distribution.
  - $O(1)$ links per node, $O(\log^2 n)$ routing, if $\log n$ long range links $O(\log n)$ routing

- combines several ideas:
  - DHT
  - Kleinberg's Small World model
  - other improvements
SYMPHONY: A KLEINBERG BASED DHT
SYMPHONY: A KLEINBERG BASED DHT

Each nodes defines:

- a link with its predecessor and one with its successor
- possibly some links with close neighbours (in the figure a link with the successor of its successor)
- $k$ ($k \geq 1$) long distance links
Long range links are defined in the following way:

- choose a number $x$ according to an **harmonic distribution**

\[
p_n(x) = \begin{cases} 
\frac{1}{n} & \text{if } x \in \left[\frac{1}{n}, 1\right] \\
\frac{1}{x \cdot \ln n} & \text{otherwise} \\
0 & \text{otherwise}
\end{cases}
\]

- $x$ is the distance of the target of the long range link from the node source of the link
- detect the point $y$ distant $x$ (clockwise) with respect to itself
- contacts the peer which manages of $y$ and tries to define a long range link with $y$
Harmonic probability probability distribution:

\[
p_n(x) = \begin{cases} 
\frac{1}{x \cdot \ln n} & \text{if } x \in \left[\frac{1}{n}, 1\right] \\
0 & \text{otherwise}
\end{cases}
\]

- \(n\) is the number of nodes in the ring
- \(x \in [1/n, 1]\) is the ring distance of the target from the source of the long range link
- in the harmonic distribution
  - the probability of a long range link depends on the inverse of the distance (Kleinberg) and also from the number of nodes
  - given a distance, if the number of nodes increases, the probability logarithmically decreases: more nodes on the ring, more close links
In Chord the computation of the long range links is deterministic and complex.

Symphony exploits a statistical approximation to obtain a similar result, but in a more efficient way.

- lower number of messages
Symphony: A Kleinberg Based DHT

- Greedy Routing Algorithm (inspired by Kleinberg):
  when a node $n$ looks for a key, it sends the key through the link (short or long range) which minimizes the distance, computed clockwise, between the target and the key.

- Theorem: consider a Symphony DHT with $n$ nodes, where each node has $k$ long range links. The average number of nodes which has to be contacted before reaching the node managing a key is $O(\log^2 n / k)$.

- If $k = \log n$ the routing requires $O(\log n)$ hops, like Chord.

- A sort of probabilistic Chord.

- The result is valid only for the harmonic distribution. If, we choose a uniform distribution, the complexity grows as the radix of $n$. 

• The value $K$ is a upper limit to the number of connections managed by each node: it may be defined at configuration time.

• A node chosen as target of a long range link, may refuse the connection, if it has already exceeded the limit of $K$ opened connections.

• In this case, the node that has required the connection (the one which wanted to define the long range link), determines a new value of $x$, by applying the harmonic distribution.

• Symphony also checks that multiple links are not defined between the same pair of nodes.
to evaluate the probability distribution, each node needs to know \( n \), the total number of nodes on the overlay

the computation of the exact number of nodes, in a distributed setting is not simple. Exploiting a gossip algorithm?

Symphony exploits a simpler strategy based on a heuristics:

- if the identifiers are uniformly distributed on the ring, then each node approximatively manages a segment whose length is \( 1/n \).

- consider \( X_s \), the sum of the lengths of the segments managed by the \( s \) nodes of the Symphony ring, \( s << n \)

- by the knowledge of \( X_s \) it is possible to approximate the number of nodes in the whole ring
• let us consider $s$ nodes:
  • if the distribution is uniform, each node manages approximately a segment the same length
  • then, the following proportion holds
    \[ s : X_s = n : l \]
    where $L$ is the total length of the Symphony ring
  • this implies:
    \[ n = s/X_s \]

• generally, consider the predecessor, the successor node and itself

• if $s=3$ each node knows the length that it manages and the length of the segments managed by its neighbours
A new node joining to the ring:

- chooses (hash function) the identifier $ID$ in the interval $[0,1]$
- defines a contact with a bootstrap node $B$ whose address is known
- detects the node that manages $ID$, by a greedy routing
- connects to the close neighbours on the ring (local contacts)
- estimates the number of nodes on the Symphony ring through the heuristics
- connect to $k$ neighbours chosen at random (remote contacts)
  - choose a value $x \in [0,1]$ according to harmonic probability distribution:
    \[
    P(X = x) = \frac{1}{x \cdot \log n}
    \]
    where $n$ is the number of nodes in the overlay
  - try to define a remote connection with the node managing the point whose distance is $x$.
- the long range links are then periodically updated to face churn.
Voluntary leave of a node n from the ring:

- eliminate all the long range links
- for each long range link pointing to n and starting from node y, notify to y the leave of n
- each y must detect a new target for its long range link
- the neighbours of n update their short range links
- the estimate of the total number of nodes of the network is recomputed by each neighbour of n.
SYMPHONY: CONCLUSIONS

• **Greedy Routing** (“a la Kleinberg”): each request is routed toward the node that manages the segment “closest” to the key in the request.

• Theorem: the average number of steps of the Symphony routing algorithm with \( k = O(1) \) remote connections defined by each node is inversely proportional to \( k \) and proportional to \( (\log n)^2 \).