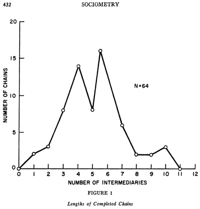
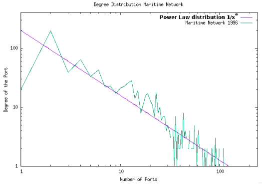
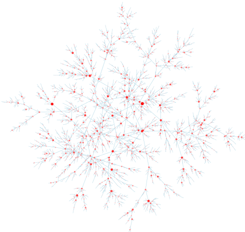
# Metrology and Measures of Complex Networks



http://litis.univ-lehavre.fr/∼guinand

# Understanding our World

## Stanley Milgram

- in 1967 a psychologist, Stanley Milgram, asked a simple question:
  **how many intermediaries are there between any two persons?**
- for instance between you and Prince of Monaco?
- if you know personaly him: 0
- if you personaly know someone who personaly knows him: 1
- if you know someone who knows another person who personaly knows him: 2...

# Understanding our World

## Stanley Milgram

- in 1967 a psychologist, Stanley Milgram, asks a simple question:
  **how many intermediaries are there between any two persons?**
- for instance between you and Prince of Monaco?
- if you know personaly him: 0
- if you personaly know someone who personaly knows him: 1
- if you know someone who knows another person who personaly knows him: 2...

we live in a **Small World**

## The Experiment

- *296 arbitrarily selected individuals in Nebraska and Boston are asked to generate acquaintance chains to a target person in Massachusetts employing "the small world method" [Milgram 1967] (..) Sixty-four chains reach the target person. Within this group the mean number of intermediaries between starters and targets is 5.2* [Travers and Milgram 1969]
- the method consists, in a graph, in choosing the next intermediary suspected to be the closest person to the target ∼ supposed shortest path to the target

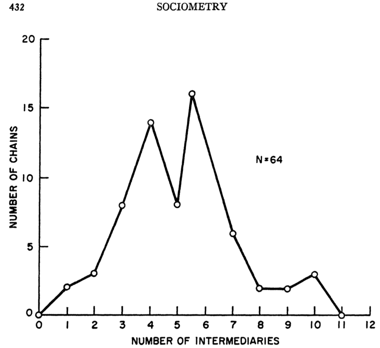



FIGURE 1

*Lengths of Completed Chains*

For a grid (Moore) of comparable size (17x17 = 289 vertices), the radius is 17 and the diameter is 34. For a Random Graph? Hypothesis on the average degree...

the term **Small World** has been kept by the scientific community up to now

# Understanding our World

## Questions

- the results obtained by Milgram were **counter intuitive** and raised **questions** about our understanding of the **nature and the structure** of the relationships between elements within real networks
- investigating scientifically this point first asks some questions in relation with our knowledge:

  Does the **underlying topology** of this social network is similar to some known graph topologies?
- if its characteristics are close to random graphs then we may suppose that the links appeared randomly and present no particular semantic
- if its characteristics are close to regular graphs (grids, torus, hypercubes, etc.) this might be a clue about a possible underlying and structured organization, what could be the cause of such an organization?
- if not, do all real networks have similar characteristics?
- if some characteristics are common to several real networks, what could be the origin of this similarities?

# Understanding our World

- but... at that point (Milgram 1967) it is difficult to compare the underlying social network with classical standard graphs (random, grids, etc) since we only have one experiment and absolutely no idea about the actual topology of the network

  $\longrightarrow$ Questions:

1. what metrics can we use for comparing graphs?
2. what are the values of these metrics for classical standard graphs?
3. How can we obtain graphs from real networks?
4. What could be the mechanisms generating graphs showing similar characteristics as the ones of real networks?

# Measures / Metrology

- Measure: the numerical value associated to a quantity or a quality of a given characteristic of a graph/node/link
- Metrology: the method for measuring the targeted quantities characteristics

## Classical Metrics

- order,
- density (number of edges / total possible number of edges),
- average degree,
- eccentricity, diameter, radius,
- chromatic number,
- etc.

## Additional Metrics

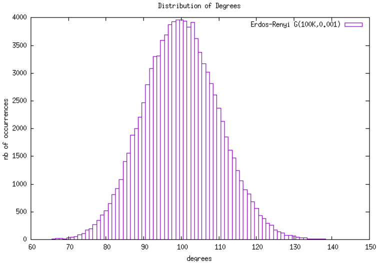
- distribution of degrees: for a given graph, what is the shape of the curve that represents the number of nodes having a given degree value
- Clustering coefficient: characterizes the fact that a vertex belongs to a group strongly connected (local version)
- Centrality: many different centrality metrics. Main purpose, highlighting nodes/edges that present particular characteristics (high degree, etc.)
- Modularity: measures the fact that the graph is composed of strongly connected nodes composing groups, themselves loosely coupled
- Assortativity: if the graph is assortative, nodes which are alike are more likely to be connected to each other (not covered by this lecture)

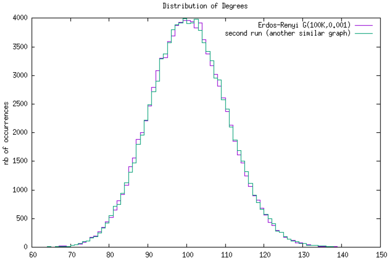
## Distribution of Degrees

### Categories

- random graphs. Example with a Erdos-Renyi graph with $n = 100000$ vertices and a probability $p = 0.001$.
- the total number of edges examined is $n(n-1)/2 \sim 5 \times 10^9$ , each created with a probability $p$, the total number of edges should be close to $5 \times 10^9 \times 10^{-3} = 5 \times 10^6$
- as each edge participates in the degree of two vertices, the average degree should be close to number of edges times 2 divided by the number of vertices:

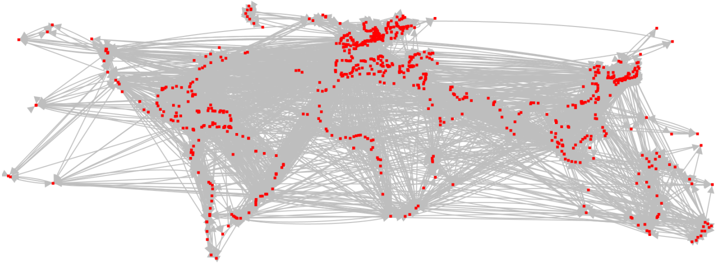
  $5 \times 10^6 \times 2/10^6 = 100$

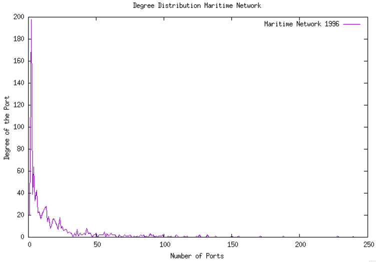



- the distribution is clearly uniform
- question: do real networks present similar degrees distributions?
- if the answer is yes then this would be a clue for considering that the growth of the network is mainly random
- if not... this would mean that underlying processes responsible of such a distribution exist and are still to be identified

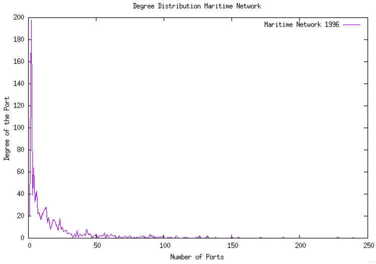
# Maritime Network

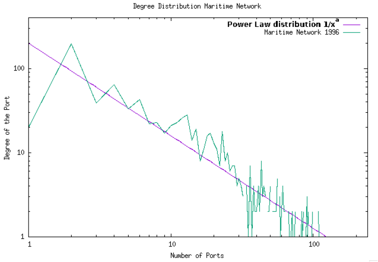





- the distribution is clearly not uniform
- a large number of ports have a small degree
- a small number of ports have a large degree

# **Power Law** degree distribution



## Clustering Coefficient

- a measure of the tendency of a node to gather with other nodes
- within real networks it has been shown that there exist groups of nodes more tightly linked within the group than with nodes outside of the group
- principle: if a node belongs to a clique and has very few links with nodes outside of the clique then its clustering coefficient should be high

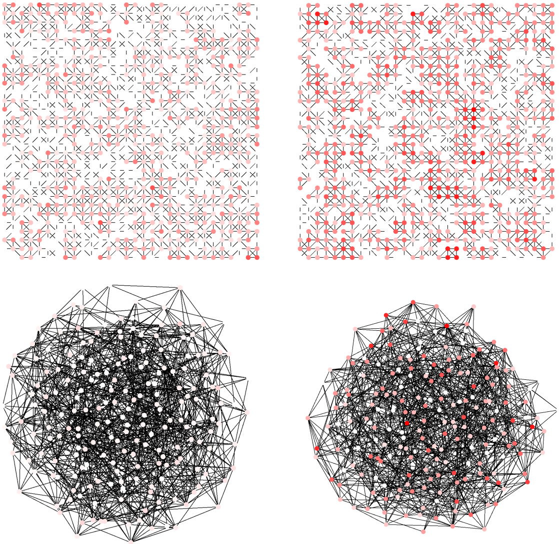
## Clustering Coefficient

- there exist two metrics: the global **clustering coefficient** and the local one[a].
- focus on the local clustering coefficient.
- given a vertex $v$ and consider its neighborhood $N_v$,
- the clustering coefficient is the density of the subgraph induced by $N_v$

[a] wikipedia.org/wiki/Clustering_coefficient

$$CC_v = \frac{2|\{u, w\} \text{ such that } u, w \in N_v|}{n_v(n_v - 1)}$$

with $n_v = |N_v|$



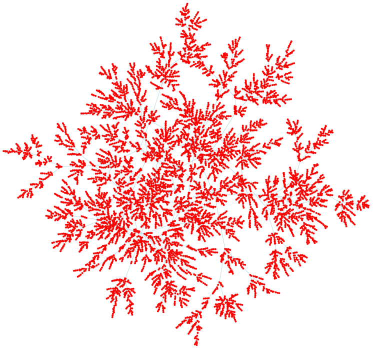
# Centrality

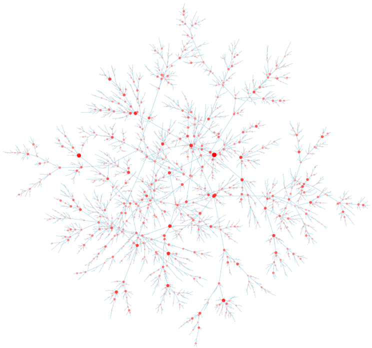
- The **centrality** metrics, measured on a vertex is supposed to identify more important vertices in the graph

  $\Longrightarrow$ there exist many ways of considering a vertex more important than another one
- in the scientific litterature we can find: *degree centrality, closeness centrality, eigenvalue centrality, betweeness centrality, harmonic centrality, Katz centrality, percolation centrality,* etc.
- we will focus of 3 of them:
  - **degree** centrality:
  - **closeness** centrality
  - **betweeness** centrality

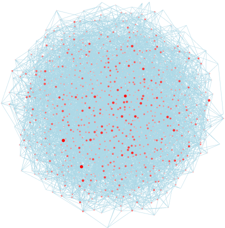
# Degree Centrality

- $C_v = \text{degree}(v)$

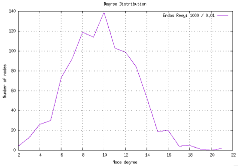


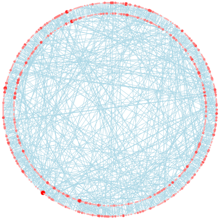


# Degree Centrality vs Degree Distribution

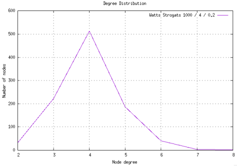


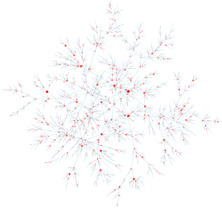
Erdös-Renyi model



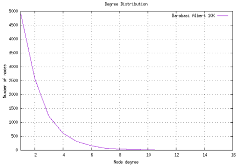



Watts Strogatz model



Barabàsi Albert model



## Closeness Centrality

- **closeness** centrality metrics attempts to capture the proximity of the node with all the other nodes of the graph (on average)
- it is then based on an average measure of the distance between this node and all the other nodes in the graph
- a distance that can be computed using Dijkstra algorithm or a similar one

$$C_v^{\text{closeness}} = \frac{(n-1)}{\sum_{u \in V} d(u,v)}$$

with $n = |V|$

## Betweeness Centrality

### Definition

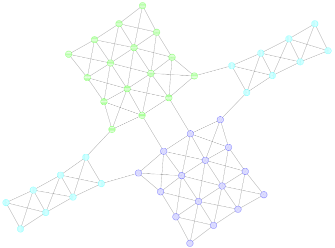
- the **betweeness** centrality measures for each node its likelihood to be present on a randomly chosen shortest path between any two vertices
- thus computing this requires counting the number of times each node belongs to a shortest path between any couple of vertices
- this can be explosive in terms of number of paths (drawing).

# Betweeness Centrality: algorithm

$G = (V, E)$
$n \leftarrow |V|$
**for** each $\{v, u\} \in E$ **do**
  $SP(u, v) \leftarrow \{$ shortest paths $v, u$ $\}$
  $n_{sp(u,v)} \leftarrow$ number of such shortest paths
  **for** each $w \in V$ s.t. $w \neq v \neq u$ **do**
    $n_w^{sp(u,v)} \leftarrow$ number of times $w \in SP(u, v)$
    $n_w^{sp} \leftarrow n_w^{sp} + n_w^{sp(u,v)} / n_{sp(u,v)}$
  **endFor**
**endFor**
**for** each $v \in V$ **do**
  $C_v^{\text{betweeness}} = n_v^{sp} / ((n-1)(n-2))$
**endFor**

## Modularity

- the modularity metrics aims at highlighting graph-topologies mainly structured as connected groups of tightly linked nodes
- typical organization: loosely coupled groups of tightly connected nodes
- a metrics close to the notion of community detection



## Modularity

### Method

- choice of the vertices defining the groups
- computation of the modularity measure

$$Q = \frac{1}{m} \sum_{\text{each group } C} \frac{m_C}{m} - \frac{\sum_{v \in V_C} d_v^2}{4m}$$

where, $G = (V, E)$ is the graph, $G_C = (V_C, E_C)$ the subgraph of group $C$, $m = |E|$, $m_C = |E_C|$

## Real Networks $\xrightarrow{?}$ graphs

- real network $\Longrightarrow$ model should be built on real data
- two situations:
  1. the data are accessible in databases, etc.
     $\longrightarrow$ generaly boring and long work (filtering/cleaning/formating) but not really difficult
  2. the data are not accessible and could not be accessed (e.g. Internet which is continously changing)
     $\longrightarrow$ strategies for discovering part of the network, generally a loop:
     observation $\rightarrow$ data $\rightarrow$ model $\rightarrow$ analysis $\rightarrow$ observation
- point 2 is still the subject of important researches
- example for case 1

## Analysis of a Complex Network: a Full Example

### Main Steps

1. find the data for building the network
2. build the network
3. measure the network
4. comparison with size-comparable classical graphs (random, grids, etc.)
5. conclusion

# Step 1: find the data

## Where to find the data?

- many datasets are now available:
  - http://networksciencebook.com/translations/en/resources/data.html
  - http://www-personal.umich.edu/∼mejn/netdata/
  - http://snap.stanford.edu/data/
  - http://vlado.fmf.uni-lj.si/pub/networks/data/
  - https://sites.google.com/a/umn.edu/social-network-analysis/resources/dataset

## Proteins Network

- we choose the Protein dataset of the networksciencebook.com web site:

  $\longrightarrow$ file `protein.edgelist.txt`

# Step 2: build the network

## raw data

- 2930 edges
- 2018 vertices

| | |
|---|---|
| 0 | 1050 |
| 1 | 229 |
| 2 | 229 |
| 3 | 467 |
| 4 | 1228 |
| 5 | 229 |
| 6 | 94 |
| 7 | 7 |
| ... | |
| 1972 | 1972 |
| 1982 | 2016 |
| 1989 | 1989 |
| 2000 | 2000 |
| 2017 | 2017 |

## Processing

- each line represent an edge under the form node src [spaces] node dest
- for our purpose we need to produce a dgs file from this data ⇒

  for each line

  ```
  0      1050
  ```

  we have to produce 3 lines:

  ```
  an 0
  an 1050
  ae 0-1050 0 1050
  ```

# Step 2: build the network

## Cleaning the data / Filtering

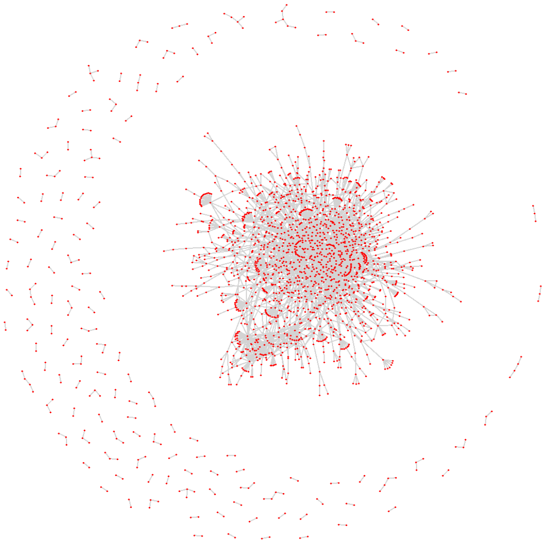
- we would like to remove all the loops
- this can be done with one line of bash command:

```
cat protein.edgelist.txt | awk 'BEGIN { RS = "\n|( *[[:digit:]]+ *)" }
      { if ($1 != $2) print $1,"-",$2 }' > proteins.txt
```

- then we can produce the dgs file:

```
cat src.txt dest.txt | sort | uniq > nodes.txt
for i in `cat nodes.txt`; do echo an $i ; done > nodes.dgs
for i in `cat proteins.txt | cut -d "-" -f 1` ; do echo $i ; done > src.txt
for i in `cat proteins.txt | cut -d "-" -f 2` ; do echo $i ; done > dest.txt
for i in `cat proteins.txt | tr ' ' '-'` ; do echo ae $i ; done > idedges.txt
paste -d " " idedges.txt src.txt dest.txt > edges.dgs
echo "DGS003" > proteins.dgs ; echo "protein 0 0" >> proteins.dgs ;
cat nodes.dgs >> proteins.dgs ; cat edges.dgs >> proteins.dgs
```

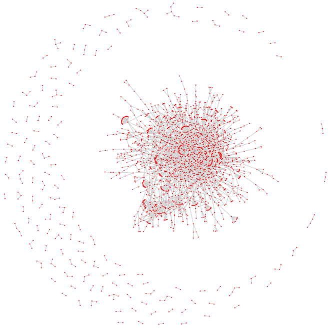
# Step 2: build the network

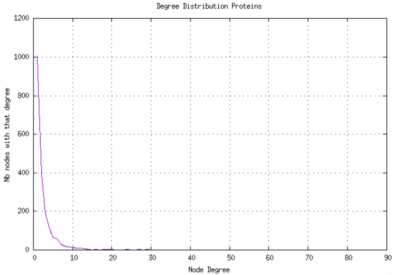


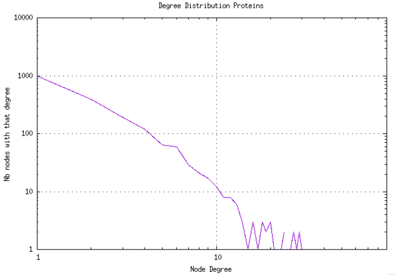
## Analysis of the model

- now we get our network as a graph, we can analyze it:
- measure some of the metrics seen so far,
- study their robustness, vulnerability
- design/study algorithms for dynamical processes like: diffusion, growth, etc.

# Metrics: degree distribution

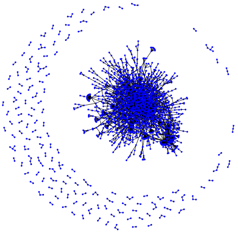




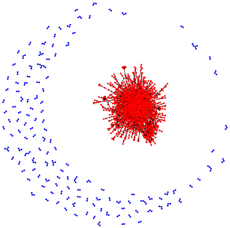



# Metrics: degree centrality

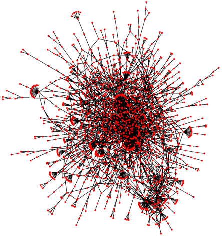
- we can observe that the graph is mainly compose of a giant connected component and of many small components
- we first "clean" the graph by keeping only the larger connected component
- this is done using graph traversal algorithm starting from a high degree vertex



original graph

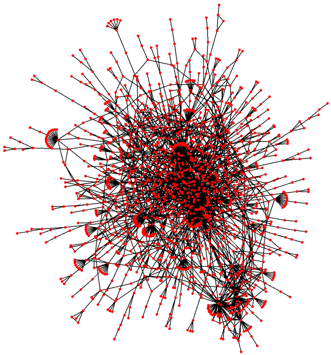


traversal of the giant component

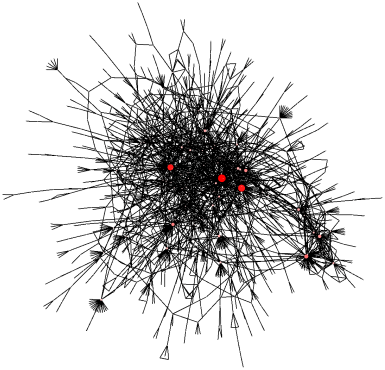


restriction to the giant component

# Metrics: degree centrality



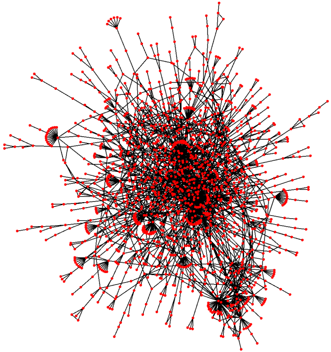
proteins graph



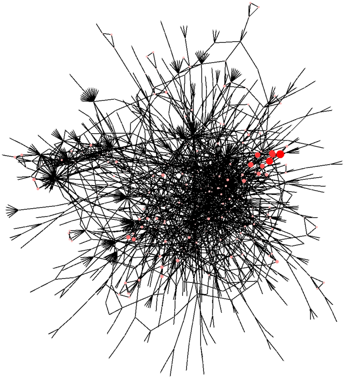
degree centrality

- very few nodes with a high degree, conform to the degree distribution of nodes

# Metrics: Clustering Coefficient



proteins graph



clustering coefficient

- clustering coefficient very low for many nodes, not surprising: $|V| = 1647$ and $|E| = 2518$
- density $\sim 2 \times 10^{-3}$, (tree: $1.2 \times 10^{-3}$) $\Rightarrow$ almost no clusters

**What could be the mechanisms generating graphs showing similar characteristics as CN?**

- the small world phenomenon
- the power law distribution of degrees

# Reproducing the Small World Phenomenon

- in 1998 Duncan Watts and Steven Strogatz proposed a method for generating graphs with the small world property

## algorithm

- starts with a lattice (in that case an "extended ring"), in which each node is connected to its $k/2$ nearest neighbors in the ring in both directions
- $m$ randomly chosen links are rewired

- if $m = 0$ the graph is regular, if $m \sim$ number of edges of the graph, the obtained graph is random, in the middle it presents the small world property

## Watts-Strogatz Model: the algorithm

build $G = (V, E)$ such that $G$ is a ring
$n \leftarrow |V|$
**for** $i \leftarrow 1$ to $n$ **do**
    **for** $j \leftarrow 2$ to $k$ **do**
        $E \leftarrow E \bigcup \{v_i, v_{(i+j)\%n}\}$
    **endFor**
**endFor**
**for** $a \leftarrow 1$ to $m$ **do**
    choose an edge and rewire it
**endFor**

# Power Law

## Generation of Power Law DD Graphs

- in 1999 Barabási and Albert proposed the preferential attachment method
- the older are more likely to be linked to many vertices

$G = (V = \emptyset, E = \emptyset)$
$V \leftarrow$ creation of node $v_0$
**for** $i \leftarrow 1$ to $n$ **do**
    choose randomly a node $u \in V$
    create node $v_i$
    add edge $\{v_i, u\}$
**endFor**