Fundamental Aspects of Decentralized Networks

Pierre Fraigniaud
CNRS
LRI, Univ. Paris Sud, Orsay, France

Summary

- Overlay networks for P2P systems
  - Semi-decentralized systems
  - Decentralized systems
    - Non-structured networks
    - Structured networks
- Large interaction networks
  - Common properties
  - Navigable networks
- Putting things together

Overlay Networks for P2P Systems

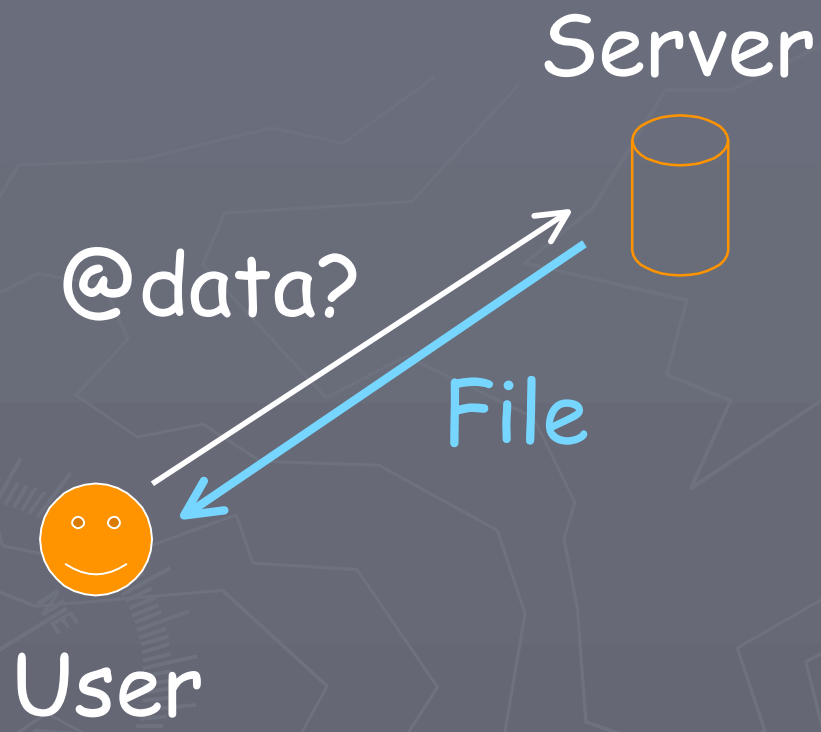
Peer-to-Peer (P2P) Paradigm

- Opposed to the master-slave paradigm
- A group of users share a common space in a decentralized manner, all playing the same role
- Objectives:
  - Share data (music, movies, etc.)
  - Share resources (computing facilities)
- Functionalities:
  - Publish
  - Search

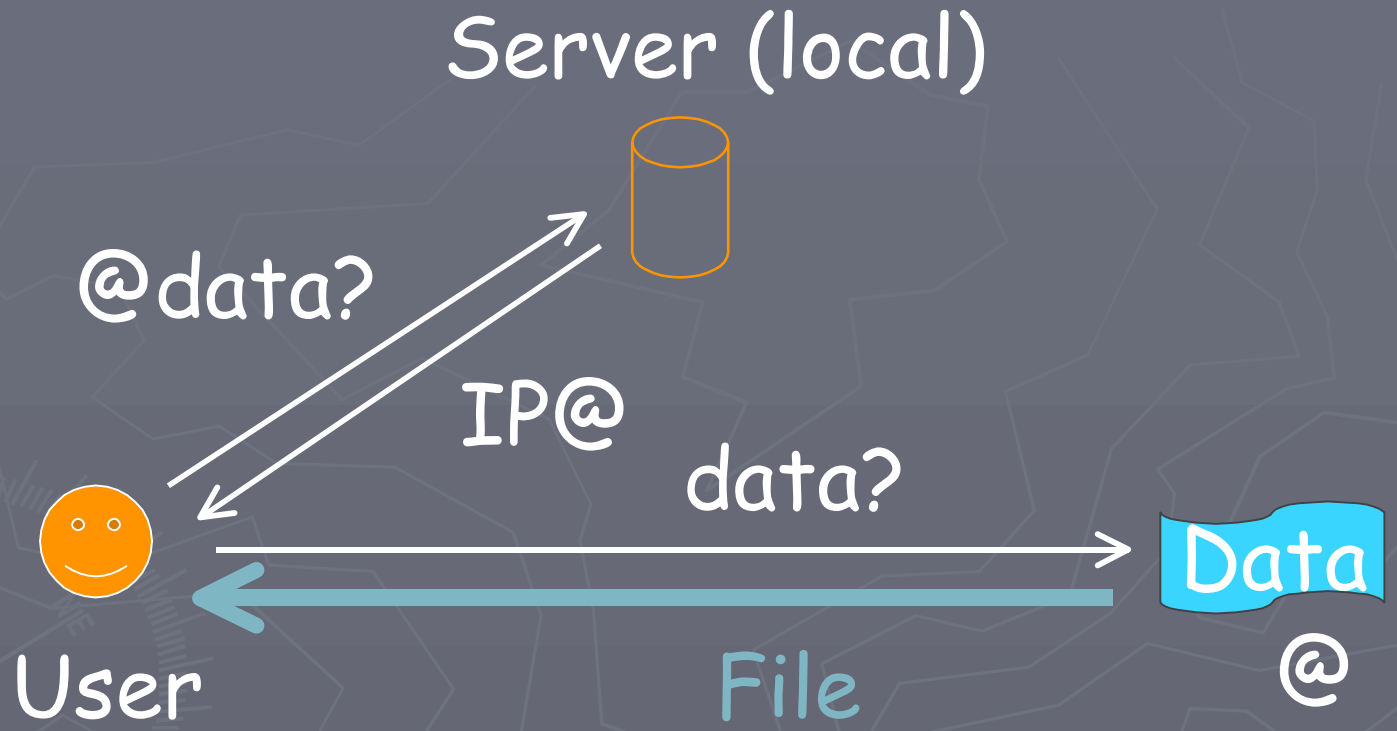
Main (Ideal) Characteristics

- No central server
- Self organization
- Users can join and leave the system at any time
- Fault-tolerance
- Anonymity (?)

Client-Server



Semi-Decentralized Systems



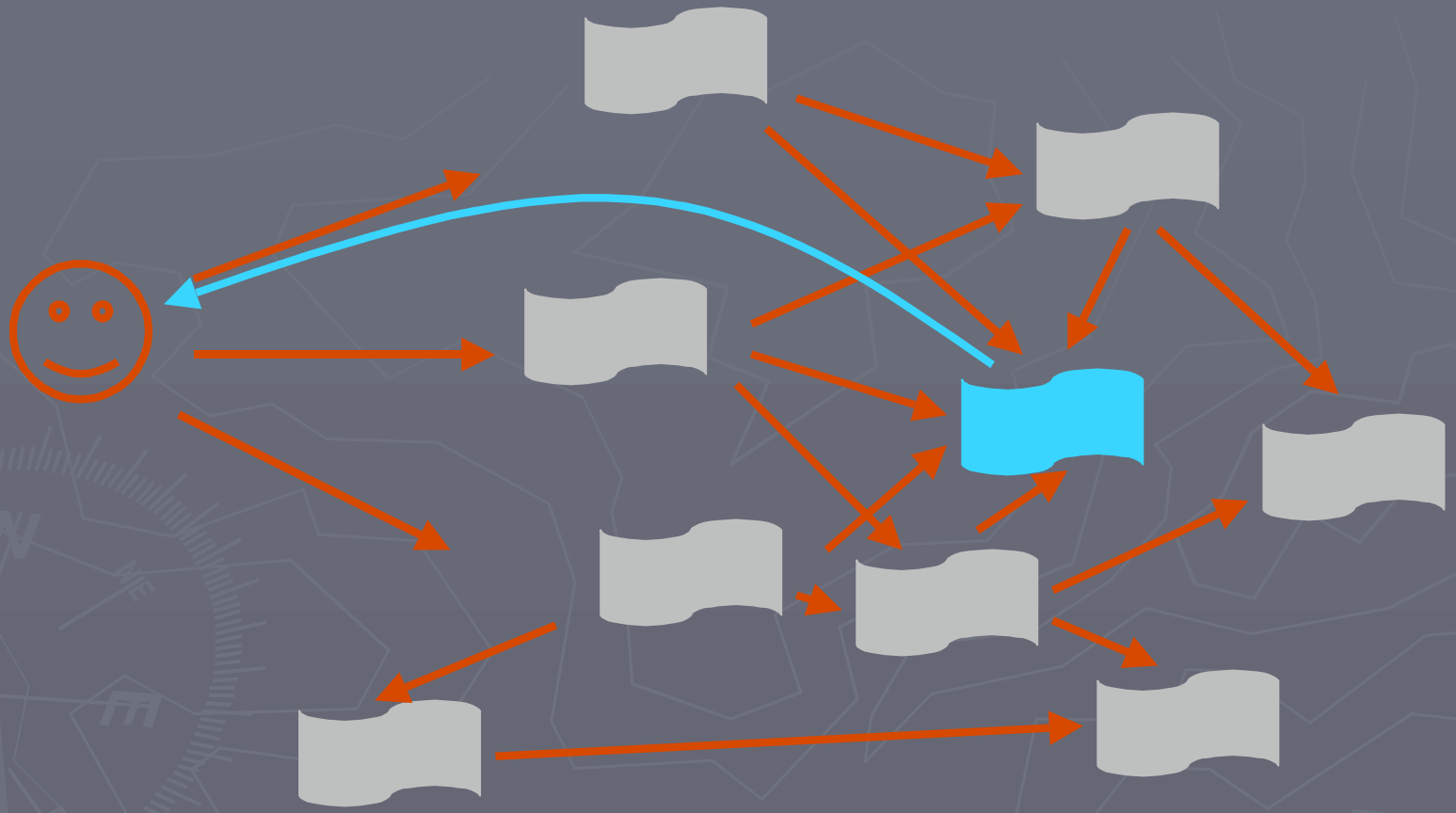
Discussion

- Pro:
  - Quick acces
  - Enable sophisticated types of request
- Con:
  - Expensive (computation and storage)
  - Bottleneck (high congestion)
  - Single point of failure

Decentralized Systems

- Nodes are connected by a logical overlay network, deployed over the Internet
- Link (u,v) means u knows the IP@ of v
- Structure of the overlay:
  - Not structured:
    - nodes connect to arbitrary nodes
  - Structured:
    - nodes are connected to specific nodes
    - in order to achieve a specific topology

Non structured networks



Discussion

- Pro:
  - Easy to implement, and cheap!
- Con:
  - High traffic load (if flooding)
  - Non exhaustive (if search is bounded)
  - Routing is hazardous

Structured networks

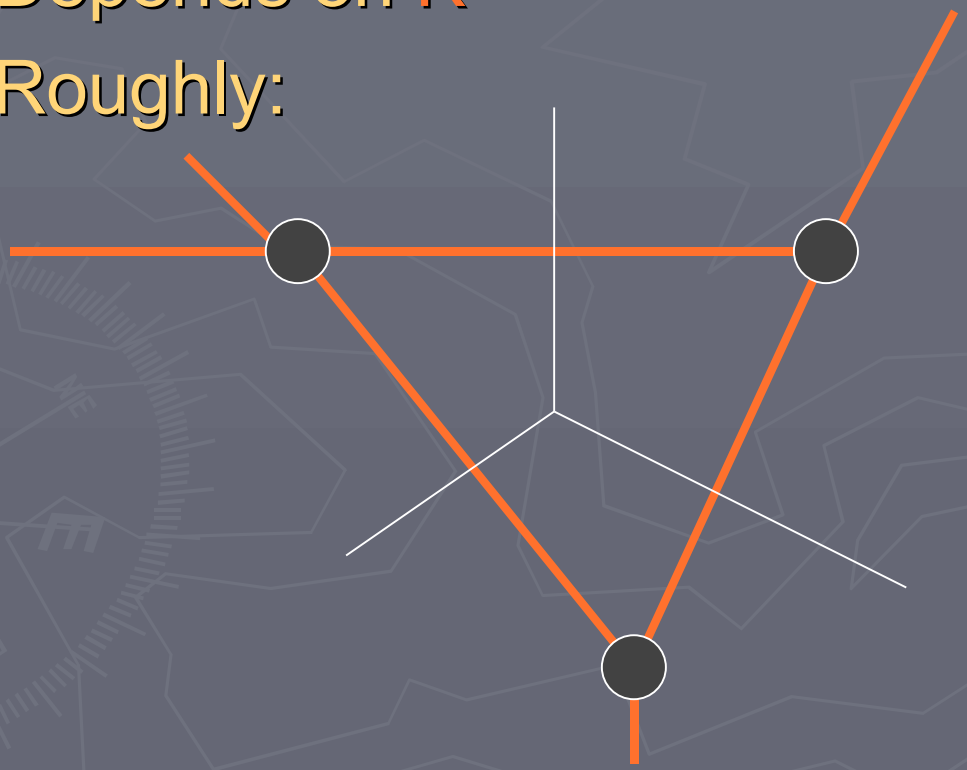
- Principles:
  - Let K be a metric space (e.g., [0,1[ )
  - Assign a label to every node

    label : { IP@ } → K
  - Assign a key to every resource

    key : { resources } → K
  - The resource r is published at the node u such that dist(label(u), key(r)) is minimal.

Principles (cont)

- Connections:
  - Depends on K
  - Roughly:



Routing

- Key-based routing (Content Addressable Networks)
- Greedy routing to κ at current node u:
  - N(u) = { neighbors of u }
  - Let v be a node such that:

    $dist(label(v), \kappa) = \min_{w \in N(u)} dist(label(w), \kappa)$
  - Route to v

Resource publication

- Node u aims at publishing resource r
  - Node u computes κ = key(r)
  - Node u tells to the node v in charge of κ that it is storing r
  - Node v stores the IP@ of u in its lookup table
- The second phase is based on the routing procedure

Searching for resources

- Node u search for resource r
  - Node u computes κ = key(r)
  - Node u asks the IP@s of nodes storing r to the node v where κ is published
  - Node v sends these IP@s to u
- The second phase is based on the routing procedure

Dynamics

- Node u leaves:
  - Reallocation of keys to u's neighbors
  - Update connections between u's neighbors
- Node u joins:
  - Connection to an entry point
  - Label computation (hash function: label(u)=hash(IP@(u))
  - Setting of u's connections
  - Reallocation of keys from u's neighbors

Discussion

- Structured networks are based on the Distributed Hash Table (DHT) paradigm
- Pro:
  - Fully distributed
  - Low traffic and load balancing
  - Exhaustive search
- Con:
  - The dynamics is a bit complex
  - Basic requests (key-based)

Problem

- Design a dynamic network (i.e., nodes can join and leave at their convenience) in which routing and updating are efficient.
- Many solutions, based on standard static graph topologies

Constraints

- Fast updates

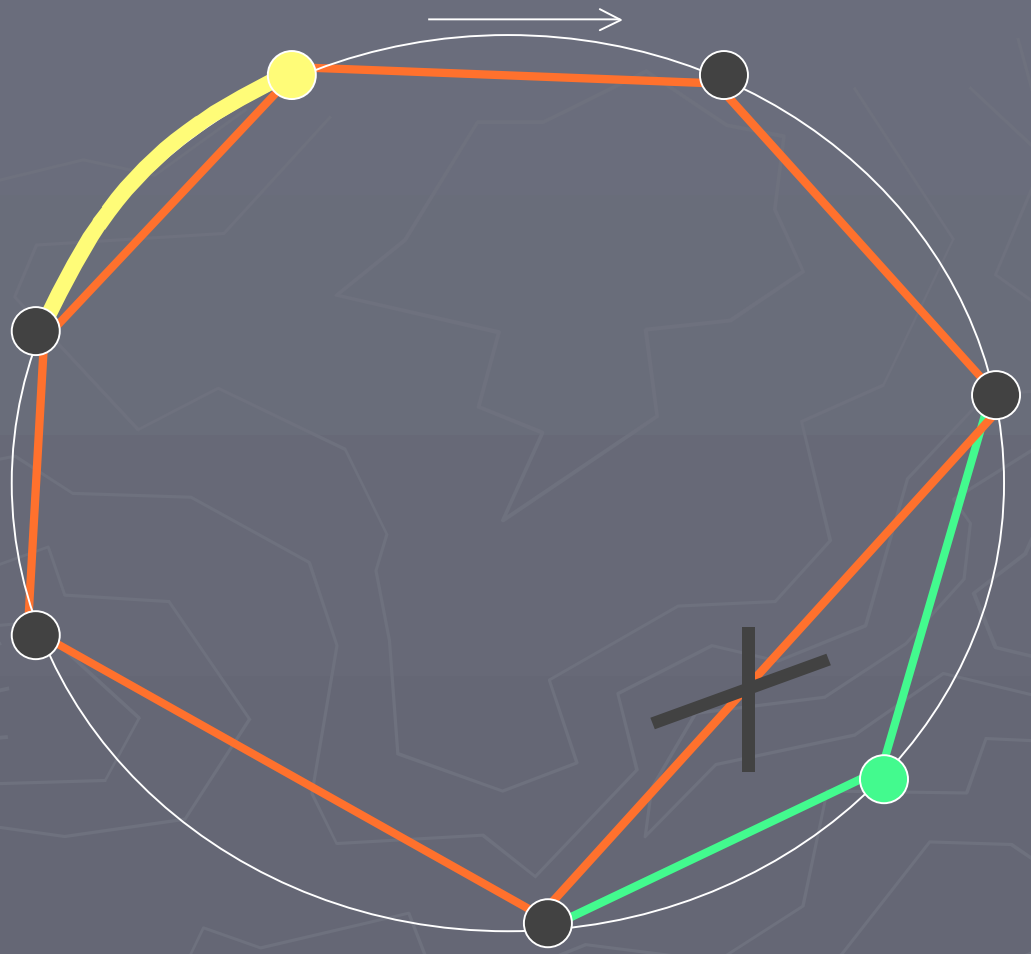
  Limited amount of control messages

  $\Rightarrow$ small degree
- Fast lookups

  Short routes $\Rightarrow$ small diameter
- Balanced routing traffic

  No hot spot

Example: the oriented ring

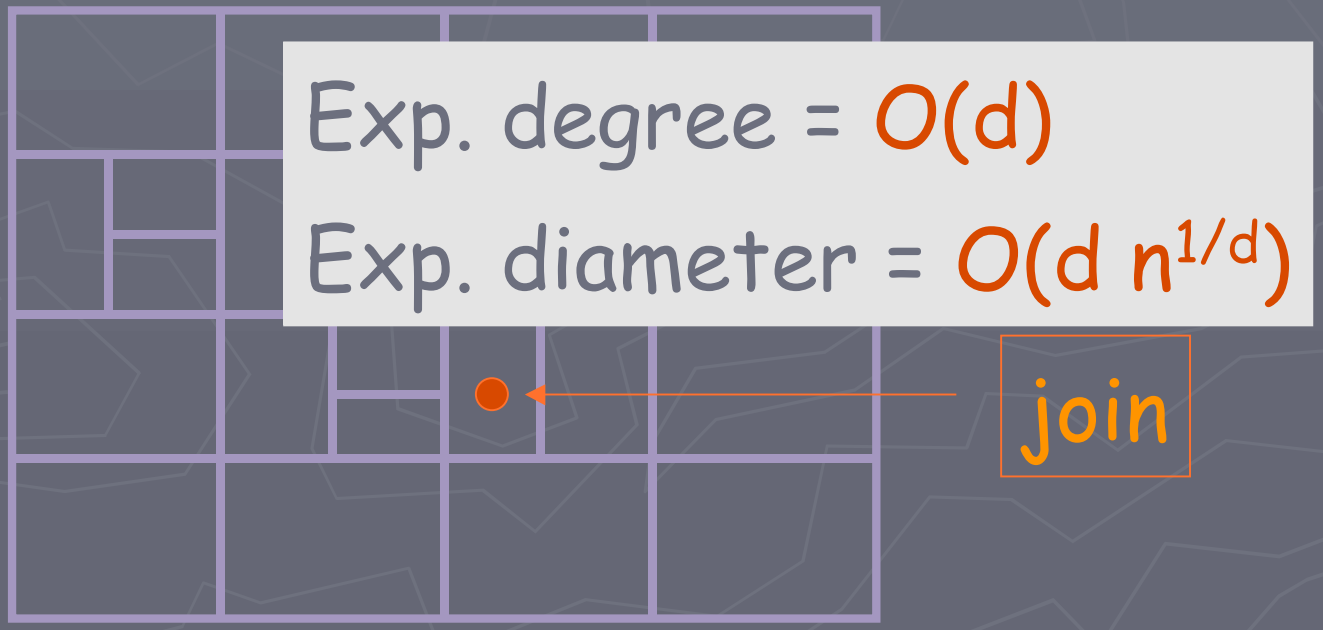


CAN

"Content-Addressable Network"

Ratnasamy,Francis, Handley, Karp, Shenker [SIGCOMM '01]

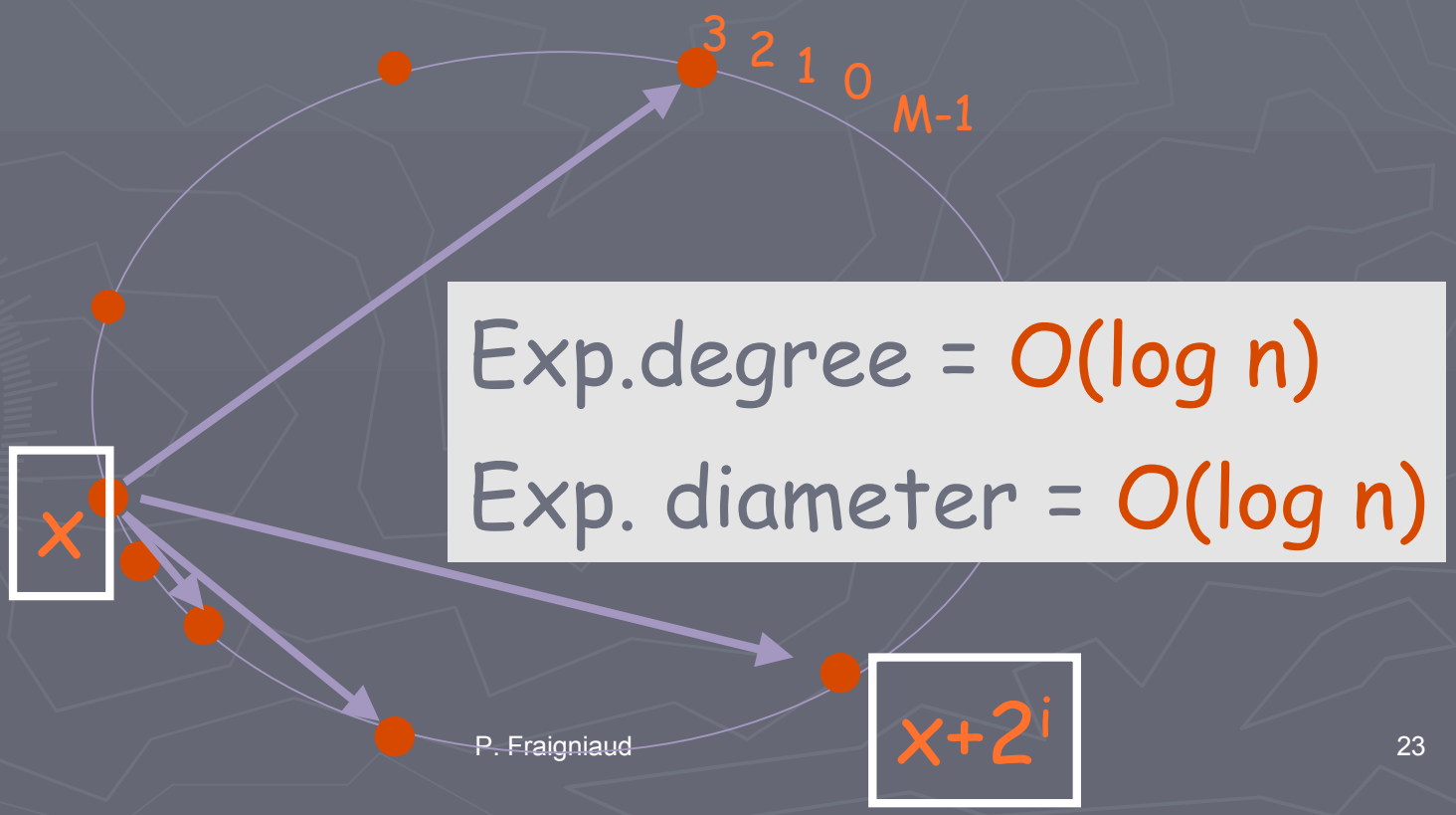
d-dimensionnal torus



Chord

Stoica, Morris, Karger, Kaashoek, Balakrishnan [SIGCOMM '01]

d–dimensional hypercube



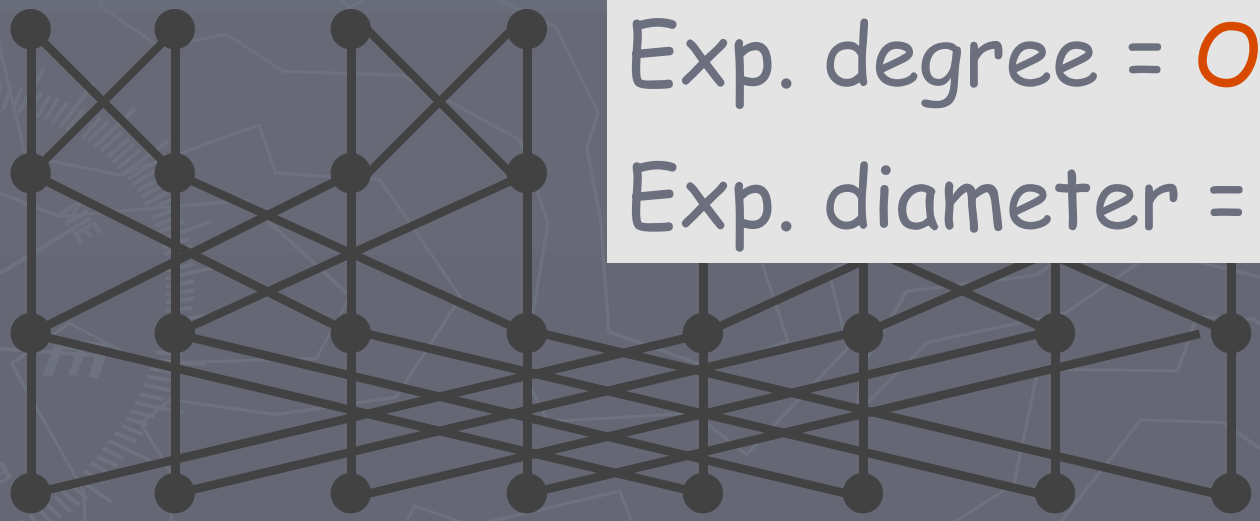
Exp.degree = O(log n)

Exp. diameter = O(log n)

Viceroy

Malkhi, Naor, Ratajczak [PODC '02]

Butterfly Network



Exp. degree = $O(1)$

Exp. diameter = $O(\log n)$

de Bruijn-based DHTs

- I. Abraham, B. Awerbuch, Y. Azar, Y. Bartal, D. Malkhi, E. Pavlov: *A generic scheme for building overlay networks in adversial scenarios*
- P. Fraigniaud, Ph. Gauron: *D2B: a de Bruijn Based Content-Addressable Network*
- F. Kaashoek, D. Karger: *Koorde: a simple degree-optimal distributed hash table*
- M. Naor, U. Wieder: *Novel architectures for P2P applications: the continuous-discrete approach*

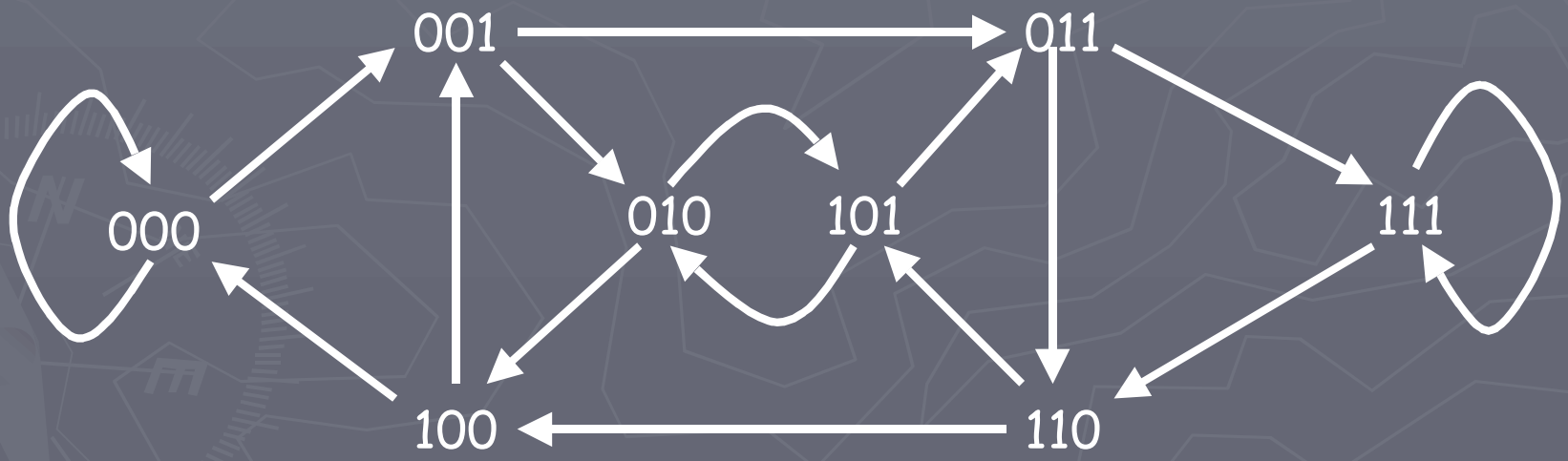
D2B

- Based on the de Bruijn graph
- Measures:
  - #key per node
  - Degree
  - Length of the routes
  - Congestion
- Performances
  - In expectation
  - With high probability (Prob ≥ 1-1/n)

De Bruijn graph

V = {binary sequences of length k}

$E = \{(x_1 x_2 \ldots x_k) \rightarrow (x_2 \ldots x_k y),\ y=0 \text{ or } 1\}$



Node and key labels

- Label = binary sequence of length $\leq m$.
- Key = binary sequence of length $= m$.

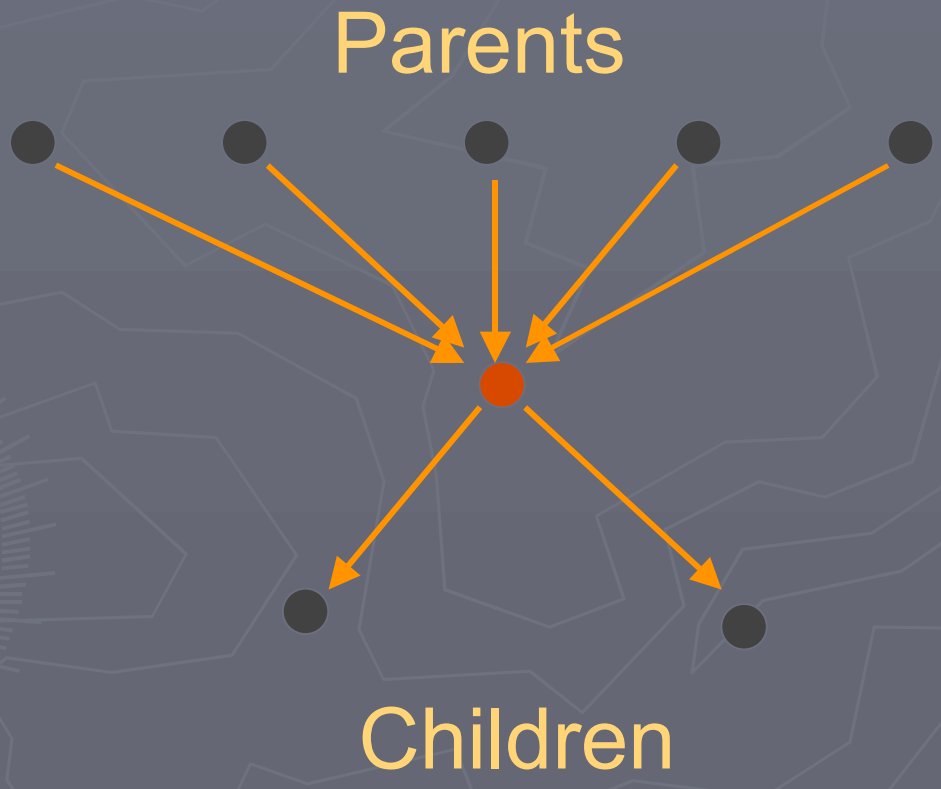
  $\Rightarrow$ up to 2^m labels and keys

  In practice, set $m=128$ or even 256
- The key κ is stored by node x if and only if x is a prefix of κ.

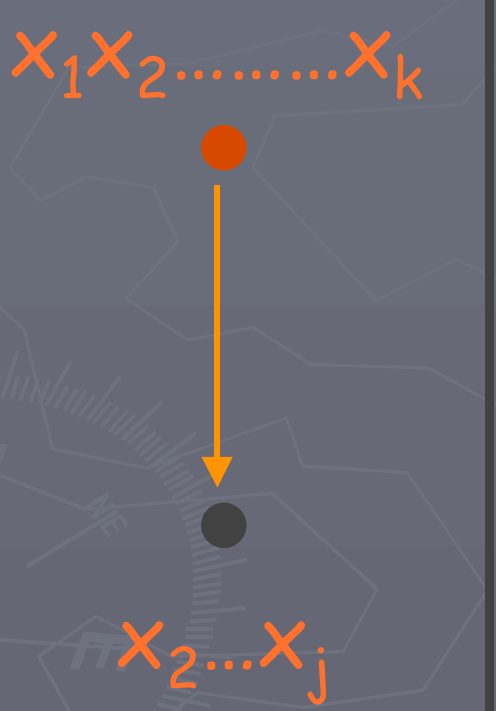
Universal Prefix Set

- Let W_i, $i=1,\ldots,q$, be q binary sequences.
- The set $S=\{W_1,W_2,\ldots,W_q\}$ is a universal prefix set if and only if, for any infinite binary sequence B, there is one and only one W_i which is a prefix of B.
- Example: {0,11,100,1010,10110,10111}
- Remark: $\{\varepsilon\}$ where ε is the empty sequence is a universal prefix set.
- By construction, the set of nodes in D2B is a universal prefix set.

Routing Connections



Children Connections and Routing



$x_2.....x_k y_1 y_2 ... y_j$

The set $\{y_1 y_2 ... y_j\}$ is a UPS

Join Procedure (1/3)

- A joining node u contacts an entry point v in the network
- Node u selects an m-bit binary sequence L at random: its preliminary label
- A request for join is routed from v to the node w that is in charge of key L

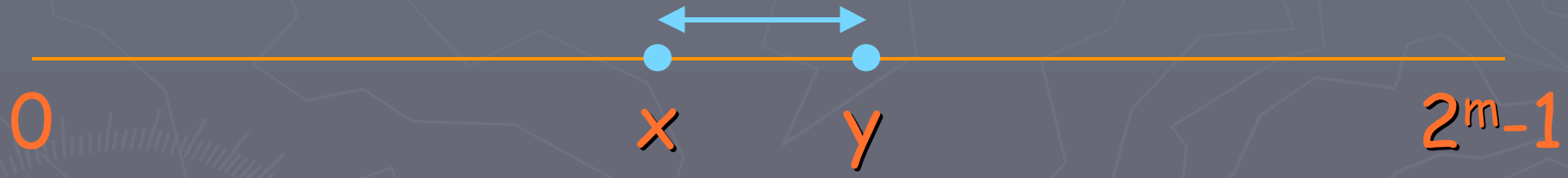
Join Procedure (2/3)

- Node w labeled $x_1x_2\ldots\ldots x_k$ extends its label to $x_1x_2\ldots\ldots x_k0$
- Node u takes label $x_1x_2\ldots\ldots x_k1$
- Node w transfers to u all keys κ such that $x_1x_2\ldots\ldots x_k1$ is prefix of κ

Join Procedure (3/3)

#keys per node (1/2)

$x_1x_2...x_k \Rightarrow x_1x_2...x_k0.....0$



#keys per node (2/2)

- Divide K in n/(c log n) intervals, each containing c log n |K|/n keys.
- Let X = #nodes in interval J starting at x
- n Bernouilli trials with probability p = c log n/n
- Chernoff bound:

$$\textbf{Prob}(|\sum X_i - np| > k) < 2\, e^{-k^2/3np}$$

$\Rightarrow$ Prob($|X - c \log n| > (3c)^{1/2} \log n$) < 2/n

$\Rightarrow$ W.h.p., there is at least one node in J

$\Rightarrow$ W.h.p., a given node manages O(|K|log n/n) keys

Lookup routing

Node $x_1x_2.....x_k$ looks for key $\kappa_1\kappa_2.....\kappa_m$

$\Rightarrow x_2.....x_k\ \kappa_1...\kappa_h$

$\Rightarrow x_3.....x_k\ \kappa_1...\kappa_h\ \kappa_{h+1}.....\kappa_{h+r}$

$\Rightarrow x_4.....x_k\ \kappa_1...\kappa_h\ \kappa_{h+1}.....\kappa_{h+i}$

$\Rightarrow x_5.....x_k\ \kappa_1...\kappa_h\ \kappa_{h+1}.....\kappa_{h+i}\ \kappa_{h+i+1}...\kappa_{h+i+s}$

$\Rightarrow x_6...x_t$

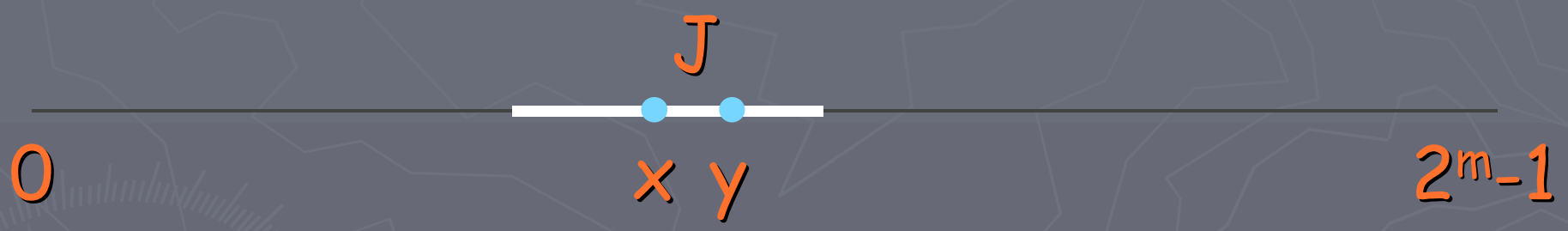
$\Rightarrow x_7...x_t\ \kappa_1.....\kappa_d$

At most k hops to reach the node in charge of the key

$\kappa_1\kappa_2.....\kappa_m$

Length of node label (1/2)

$x_1x_2...x_k \Rightarrow x_1x_2...x_k0.....0$



$|J|=c\ |K|\ \log n/n$

Length of node-label (2/2)

$\text{Prob}(|X - c \log n| > (3c)^{1/2} \log n) < 2/n$

$\Rightarrow$ W.h.p., at most $O(\log n)$ nodes in J

$\Rightarrow$ x manages at least $|J| / 2^{O(\log n)}$ keys

$\Rightarrow$ $k \leq m - \log|J| + O(\log n)$

$\Rightarrow$ $k \leq O(\log n)$

$\Rightarrow$ W.h.p., a lookup route is of length $O(\log n)$

Degree and congestion

- W.h.p., degree = O(log n) using similar techniques (expected degree O(1))
- Congestion = proba that a node is traversed by a lookup from a random node to a random key = O(log n/n)

Summary: Expected properties

	Update	Lookup	Congestion
CAN	O(d)	$O(dn^{1/d})$	$O(d/n^{1-1/d})$
Chord	O(log n)	O(log n)	O(log n/n)
Viceroy	O(1)	O(log n)	O(log n/n)
D2B	O(1)	O(log n)	O(log n/n)

Large Interaction Networks

Interaction Networks

- Communication networks
  - Internet
  - Ad hoc and sensor networks
- Societal networks
  - The Web
  - P2P networks (the unstructured ones)
- Social network
  - Acquaintance
  - Mail exchanges
- Biology, linguistics, etc.

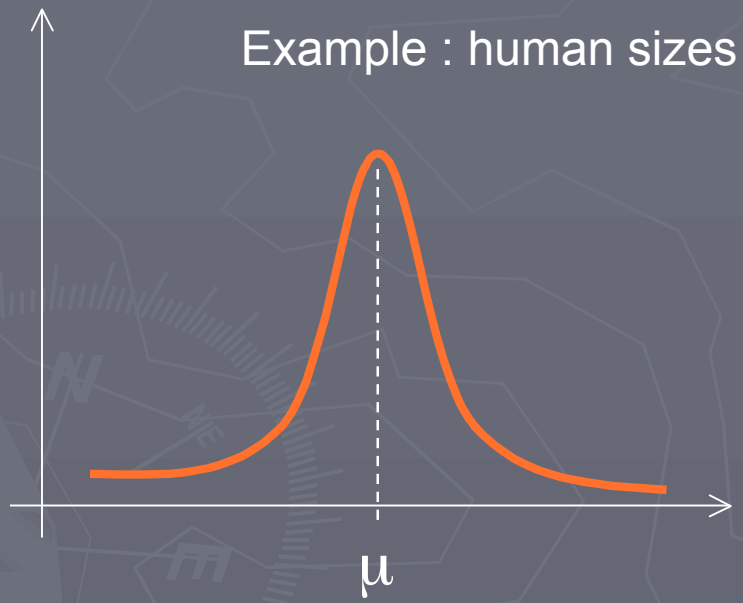
Common statistical properties

- Low density
- "Small world" properties:
  - Average distance between two nodes is small, typically O(log n)
  - The probability p that two distinct neighbors u_1 and u_2 of a same node v are neighbors is large.

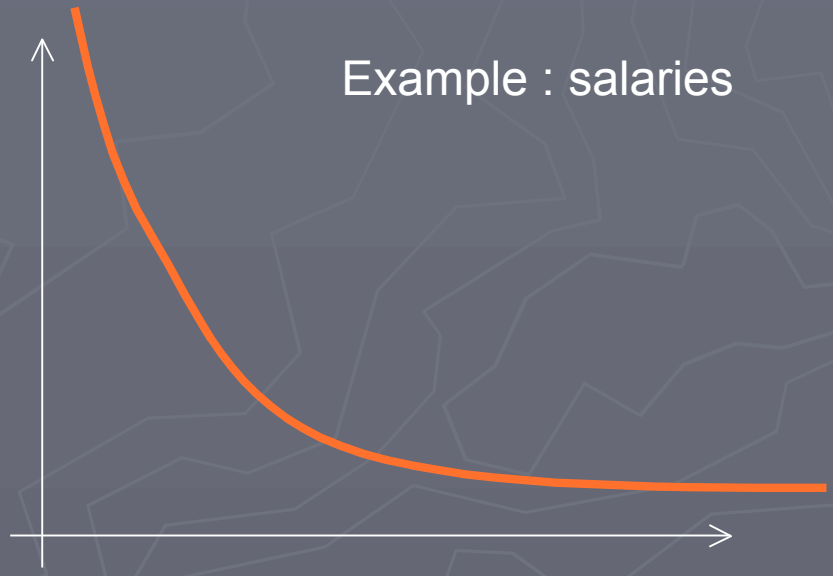
    p = clustering coefficient
- "Scale free" properties:
  - Heavy tailed probability distributions (e.g., of the degrees)

Gaussian vs. Heavy tail

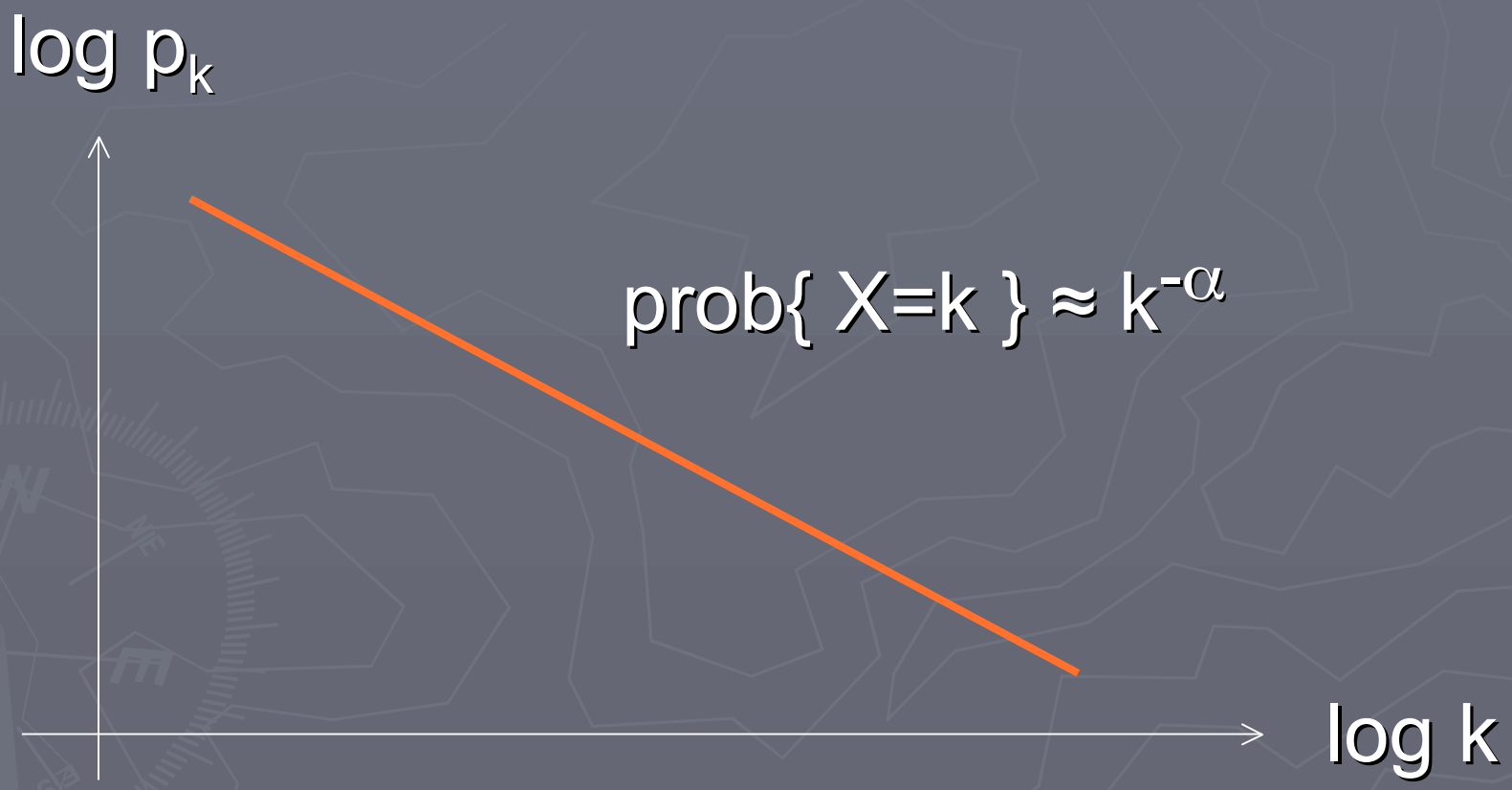
Example : human sizes



Example : salaries



Power law



Random graphs vs. Interaction networks

- Random graphs ($p \approx \log(n)/n$):
  - low clustering coefficient
  - Gaussian distribution of the degrees
- Interaction networks
  - High clustering coefficient
  - Heavy tailed distribution of the degrees

New problematic

- Why these networks share these properties?
- What model for
  - Performance analysis of these networks
  - Algorithm design for these networks
- Impact of the measures?

More insights available at:

http://www.liafa.jussieu.fr/~latapy/

Milgram Experiment

- Source person s (e.g., in Wichita)
- Target person t (e.g., in Cambridge)
  - Name, professional occupation, city of living, etc.
- Letter transmitted via a chain of individuals related on a personal basis
- Result: "six degrees of separation"

Navigability

- Jon Kleinberg (2000)
  - Why should there exist short chains of acquaintances linking together arbitrary pairs of strangers?
  - Why should arbitrary pairs of strangers be able to find short chains of acquaintances that link them together?
- In other words: how to navigate in a small worlds?

Augmented graphs H=(G,D)

- Individuals as nodes of a graph G
  - Edges of G model relations between individuals deducible from their societal positions
- A number k of "long links" are added to G at random, according to the probability distribution D
  - Long links model relations between individuals that cannot be deduced from their societal positions

Greedy Routing in augmented graphs

- Source $s \in V(G)$
- Target $t \in V(G)$
- Current node x selects among its $\deg_G(x)+k$ neighbors the closest to t in G, that is according to the distance function $\text{dist}_G()$.

- Greedy routing in augmented graphs aims at modeling the routing process performed by social entities in Milgram's experiment.

Augmented meshes
Kleinberg (2000)

d-dimensional n-node meshes

augmented with d-harmonic links

$$prob(u \rightarrow v) \approx 1/(\log(n)*dist(u,v)^d)$$

Harmonic distribution

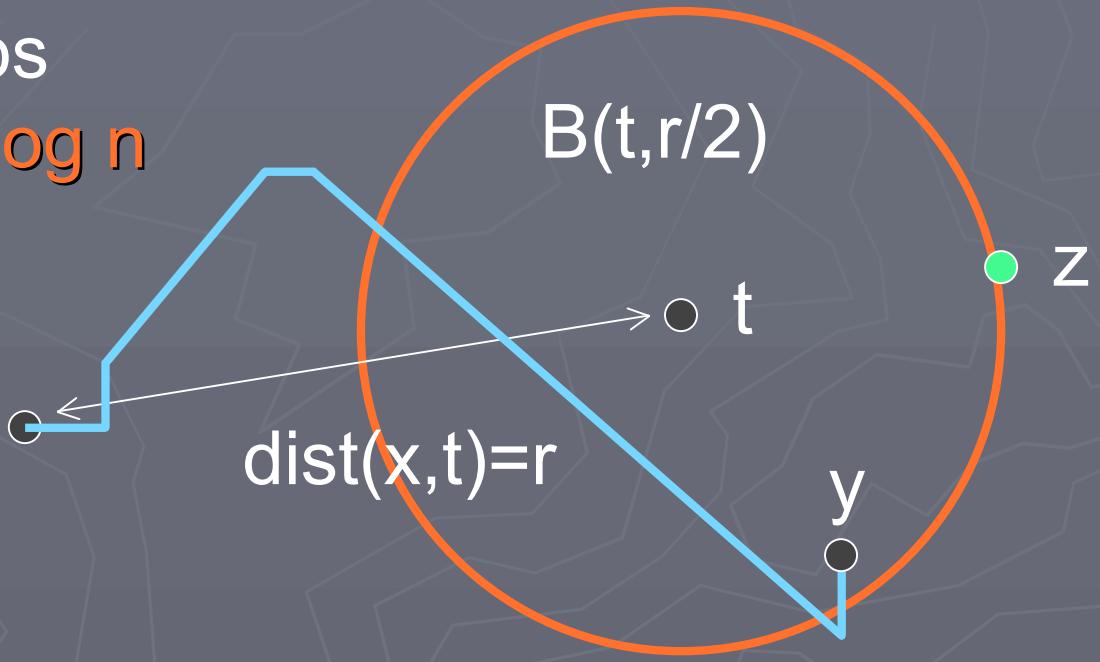
- d-dimensional mesh
- B(x,r) = ball centered at x of radius r
- S(x,r) = sphere centered at x of radius r
- In d-dimensional meshes:
  - $|B(x,r)| \approx r^d$
  - $|S(x,r)| \approx r^{d-1}$

$$\Sigma_{v \neq u}(1/dist(u,v)^d) = \Sigma_r |S(u,r)|/r^d$$

$$\approx \Sigma_r 1/r \approx \log n$$

Performances

Expected #steps
to enter B(t,r/2) = log n



For a current node x at distance r from t,

prob{x → B(t,r/2)} ≈ 1/log n

Kleinberg's theorems

- Greedy routing performs in $O(\log^2 n / k)$ expected #steps in d-dimensional meshes augmented with k links per node, chosen according to the d-harmonic distribution.
  - Note: $k = \log n \Rightarrow O(\log n)$ expect. #steps
- Greedy routing in d-dimensional meshes augmented with a h-harmonic distribution, $h \neq d$, performs in $\Omega(n^{\varepsilon})$ expected #steps.

Extensions

- Two-step greedy routing: $O(\log n / \log\log n)$
  - Coppersmith, Gamarnik, Sviridenko (2002)
    - Percolation theory
  - Manku, Naor, Wieder (2004)
    - NoN routing
- Routing with partial knowledge: $O(\log^{1+1/d} n)$
  - Martel, Nguyen (2004)
    - Non-oblivious routing
  - Fraigniaud, Gavoille, Paul (2004)
    - Oblivious routing
- Decentralized routing: $O(\log n * \log^2 \log n)$
  - Lebhar, Schabanel (2004)
    - $O(\log^2 n)$ expected #steps to find the route

Navigable graphs

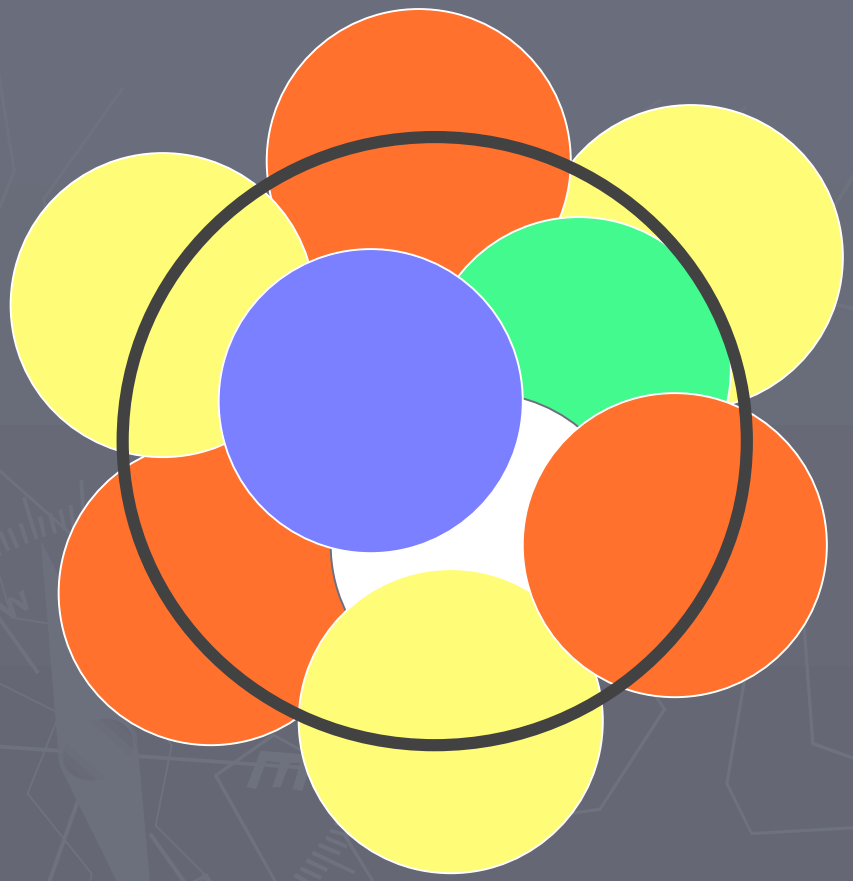
- Definition: An infinite family $\mathcal{F}$ of graphs is navigable if there exist
  - an augmentation D_G for every graph $G \in \mathcal{F}$
  - a function $f \in O(\text{polylog})$

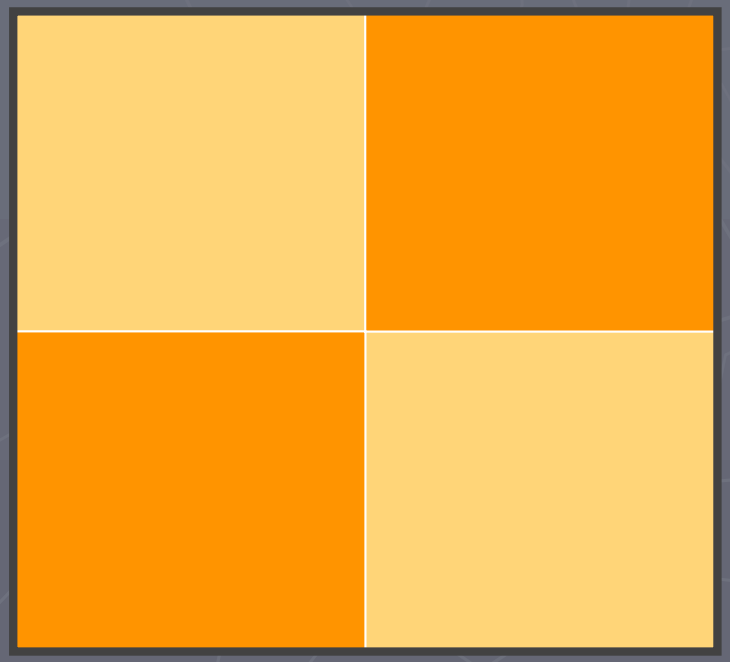
  such that, for every n-node graph $G \in \mathcal{F}$, greedy routing in (G,D_G) performs in $f(n)$ expected #steps.

Known navigable graphs

- Bounded growth graphs
  - Definition: $|B(x,2r)| \leq k\,|B(x,r)|$
  - Duchon, Hanusse, Lebhar, Schabanel (2005)
- Bounded doubling dimension
  - Definition: every $B(x,2r)$ can be covered by at most 2^d balls of radius r, $B(x_i,r)$
  - Slivkins (2005)
- Graphs of bounded treewidth
  - Fraigniaud (2005)

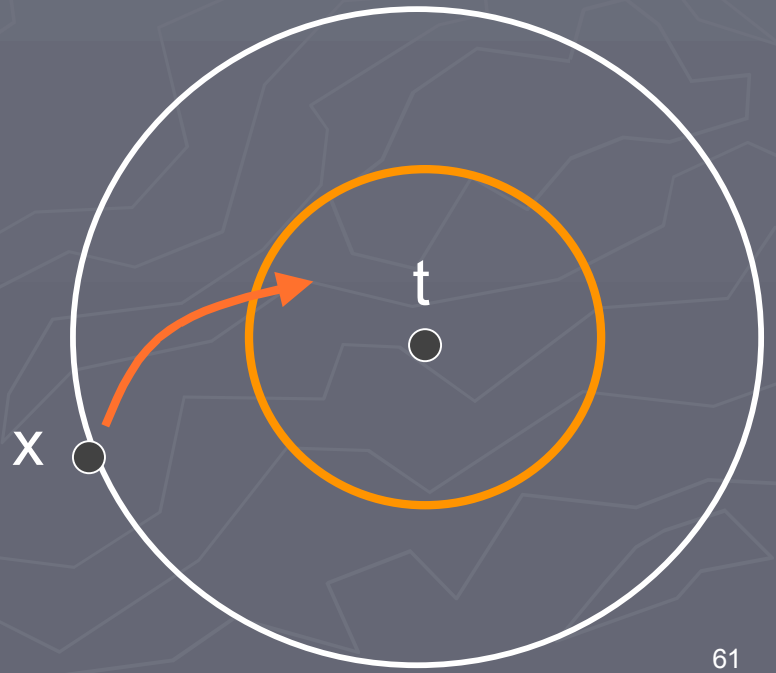
Doubling dimension





Svilkins' theorem

- Theorem: Any family of graphs with doubling dimension O(loglog n) is navigable.
- Proof: Graphs are augmented with
  - $dist_G(u,v) = r$
  - $prob(u \rightarrow v) \approx 1/|B(v,r)|$



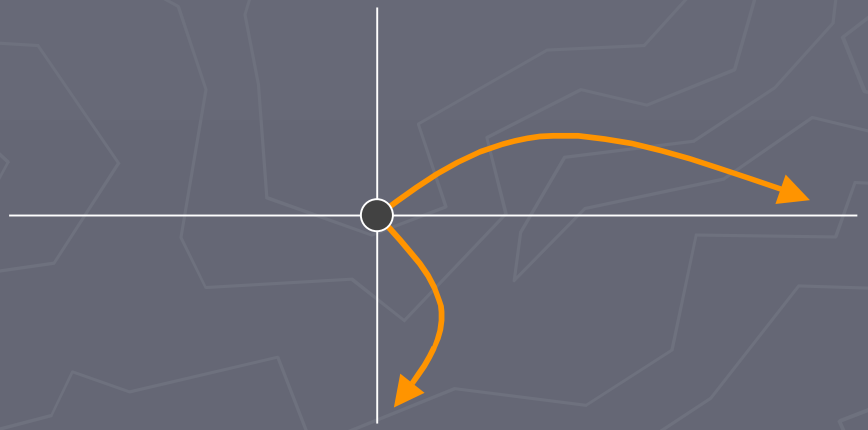
Graphs of large doubling dim.

- Remark: f(n)-dimensional n-node meshes

$$C_1 x C_2 x \ldots x C_{f(n)}$$

are navigable for any $f(n) \leq \log n$

- 1-harmonic distribution on every dimension.



But…

Theorem (Fraigniaud, Lebhar, Lotker)

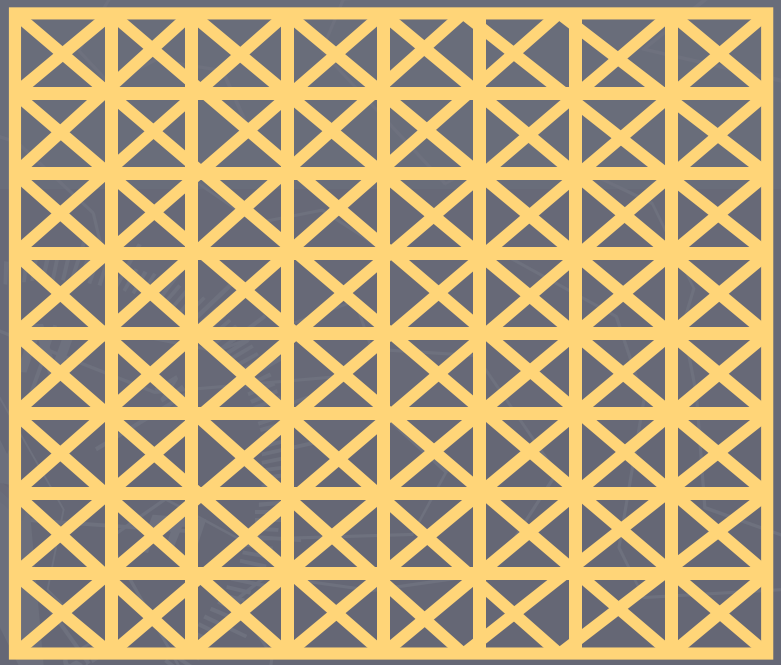
Let f such that

$$\lim \text{loglog } n / f(n) = 0$$

Any family **F** of graphs containing all graphs of doubling dimension at most $O(f)$ is not navigable.

Proof of non-navigability

- The family **F** contains the graph H_f:

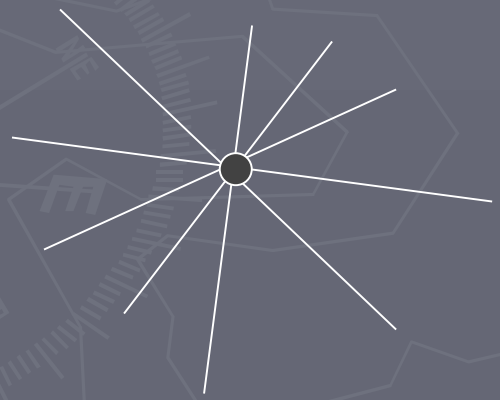


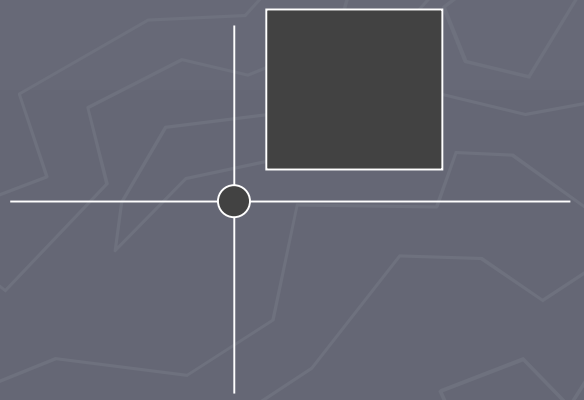
$x = x_1\, x_2 \ldots x_f$
is connected to all nodes
$y = y_1\, y_2 \ldots y_f$
such that $y_i = x_i + a_i$ where
$a_i \in \{-1,0,+1\}$

H_f has doubling dimension f

Intuitive approach

- Large doubling dimension f implies that every nodes $x \in H_f$ has choices over exponentially many directions
- The underlying metric of H_f is L_∞:





Cayley Graph

- A Cayley graph G is defined by a pair (Γ, S)
  - Γ is a group: $V(G) = \Gamma$
  - S is a generating set of Γ:

$$(u,v) \in E(G) \Leftrightarrow u^{-1}v \in S$$

- Representation of groups
- Explicit construction of expanders
- Examples:
  - $C_n = (Z_n, \{-1,+1\})$
  - $Q_d = (\{0,1\}^d, \{e_1,e_2,...,e_d\})$
  - $H_f = (C_{n^{1/f}} \times C_{n^{1/f}} \times ... \times C_{n^{1/f}}, \{g_1,g_2,...,g_{3^f}\})$

Symmetric augmentation

- An augmentation D of a Cayley graph G is symmetric if for every $g \in V$

$$prob(u \rightarrow ug) = prob(v \rightarrow vg)$$

for any pair of nodes u and v.

Symmetrization lemma

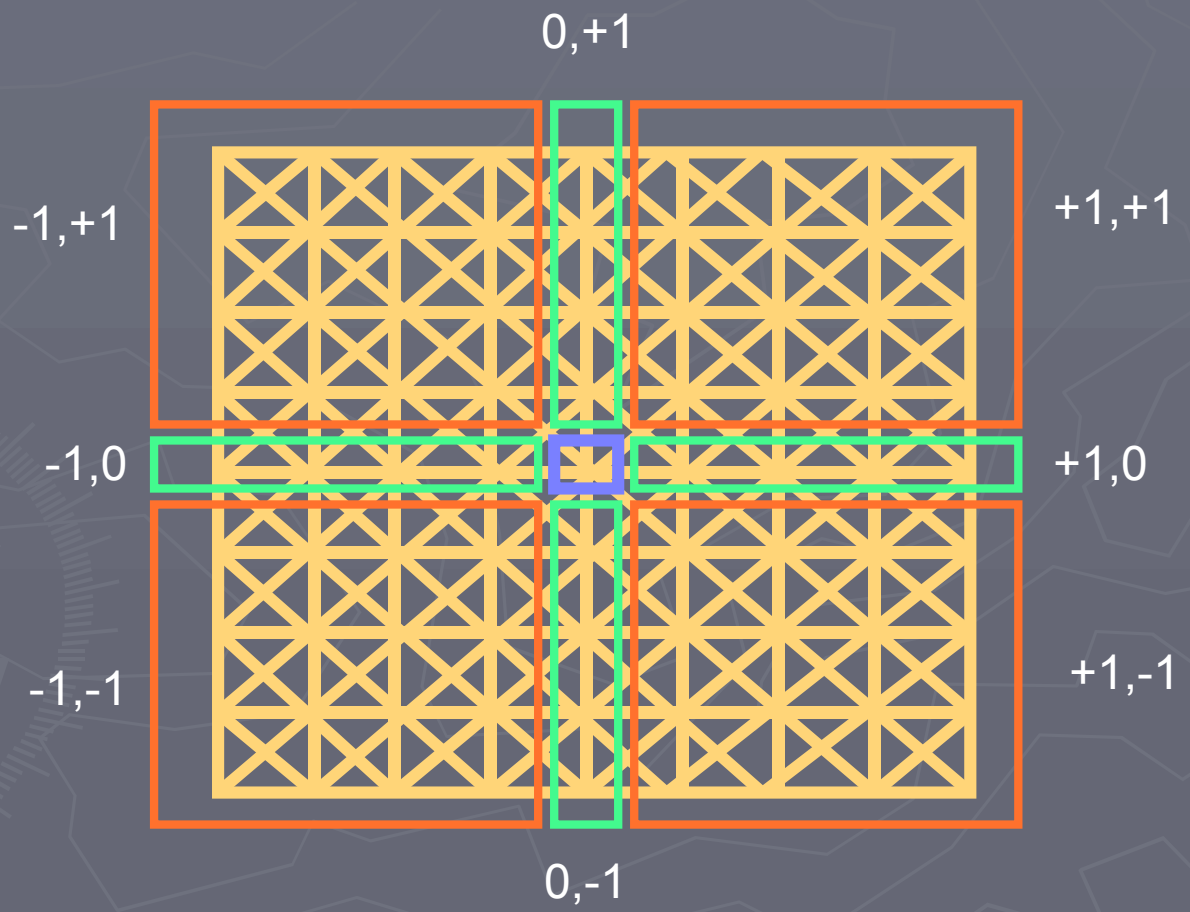
- Lemma: for any augmentation D of G, there exists a symmetric augmentation D' of G such that, for any pair s,t

  Exp(#step routing from s to t in (G,D'))

  ≤ $\max_{x,y}$ Exp(#step routing from x to y in (G,D))

- Proof:
  - x = node of G chosen uniformly at random
  - $D'_u(v) = D_{xu}(xv)$

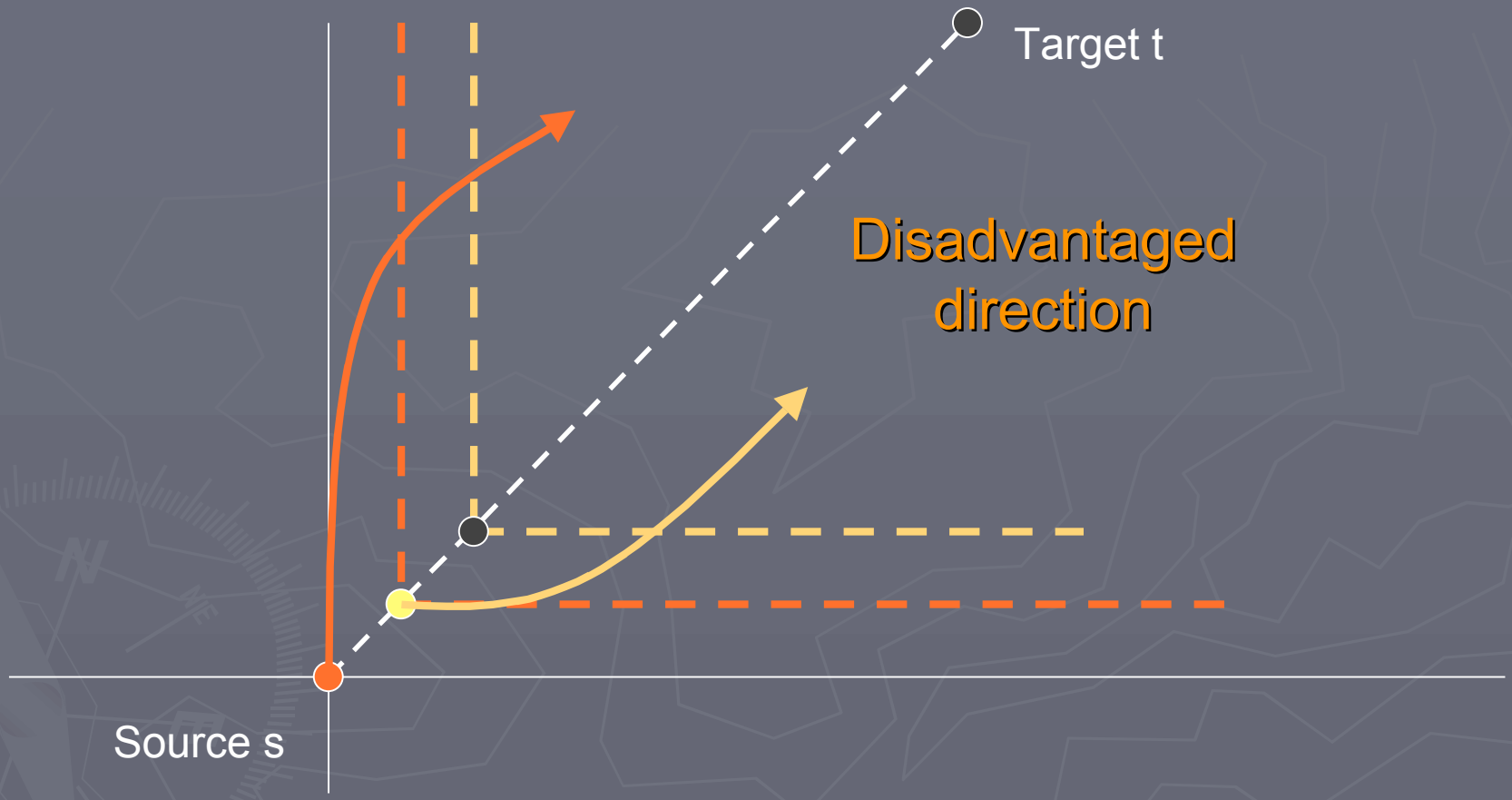
Directions



Disadvantaged directions

- $d = (d_1, d_2, \ldots, d_f) \in \{-1, 0, +1\}^f$
- $\mathrm{dir}_u(d) = \{ v = (v_1, v_2, \ldots, v_f),\ v_i = u_i + a_i\, d_i$ where $1 \le a_i \le n^{1/f}/2 \}$
- d is disadvantaged at node u if $\mathrm{prob}(u \rightarrow \mathrm{dir}_u(d)) = \min_{d'} \mathrm{prob}(u \rightarrow \mathrm{dir}_u(d'))$
- For a symmetric distribution, if d is disadvantaged at some node, then it is disadvantaged at every node.

H_f is not navigable



At every step, probability $\leq 1/2^f$ to go in the right direction

What time is it?

- **If** time **then** treewidth
- **else** next-slide

Putting things together

Using Small World and Scale Free Properties for the Design of Overlay Networks in P2P Systems

Communities

- Context: unstructured overlay networks
- Objective: create communities
- Rule: keep connected to nodes with whom you exchanged resources
- Impact: the search time is significantly reduced (observed in, e.g., Gnutella)
- Reason: acquaintances have high clustering coef., thus resources you are interested in are close to you in a network that maps the acquaintance graph.

High-degree-first search

- Context: unstructured overlay networks with power law degree distribution
- Rule: High-degree-first search strategy
  - Every node keeps track of the list of all the resources stored by its neighbors
  - DFS search visiting high degree neighbors first
- Impact: sub-linear search time
- Reason: well informed nodes are reached rapidely

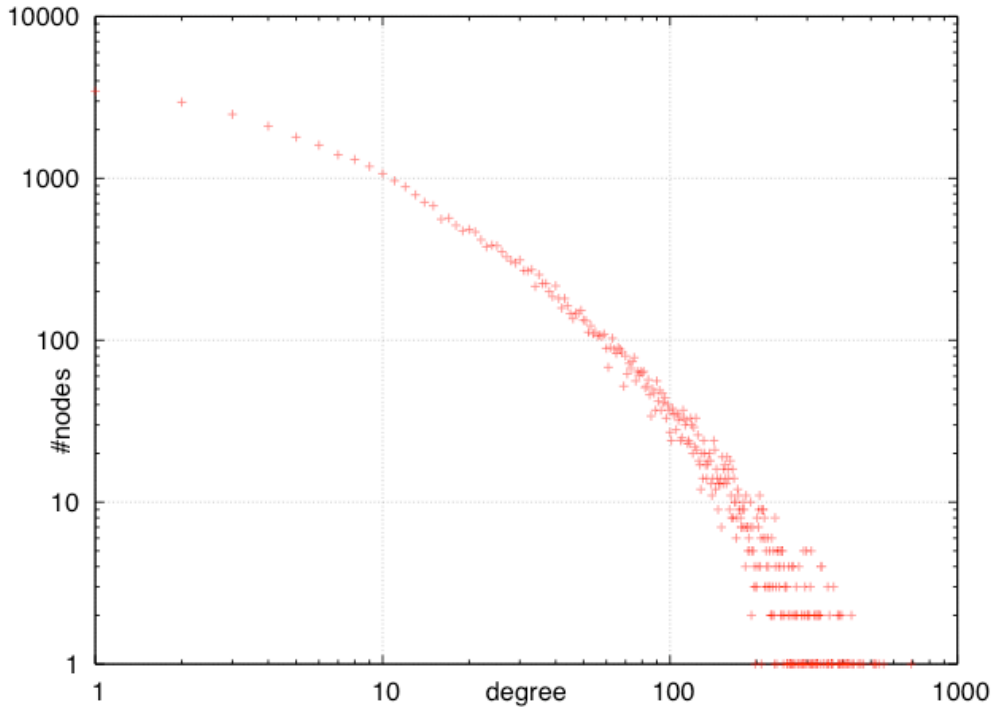
Mixing the two

- Nodes
  - join one by one, and initially connect to k arbitrary nodes
  - keep connected to nodes with whom they exchanged resources
  - store the lists of their neighbors' resources
  - perform DFS search with high-degree node priority

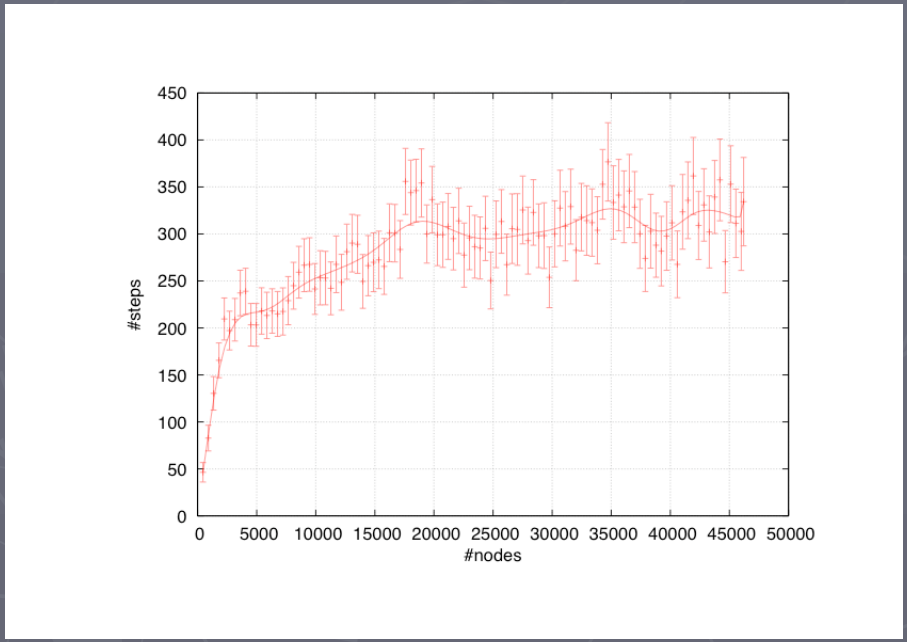
Experiments

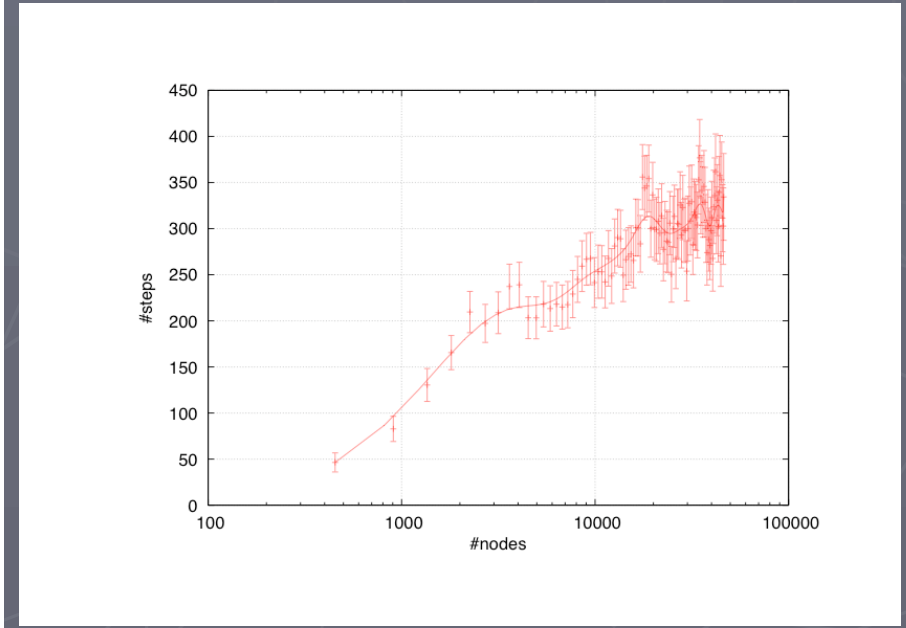
- P2P trace from eDonkey
- The trace is 2h53 long
- Involves 46,202 peers and 342,204 requests
- Simulation of each (search) request:
  - Routing from source to targets (there can be many)
  - Update connections

Degree distribution

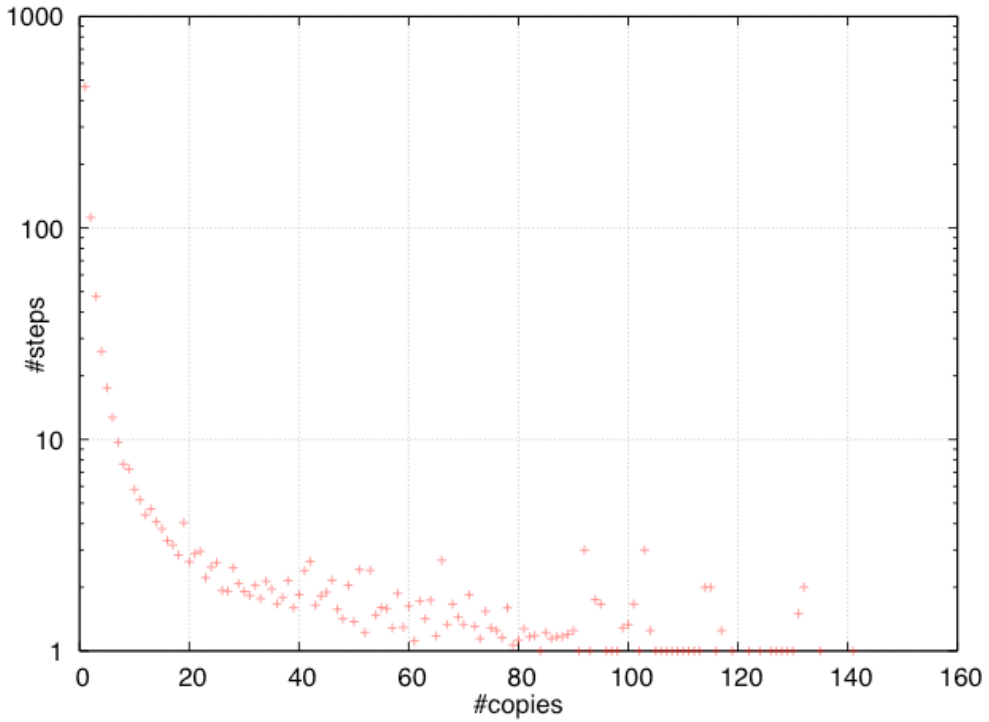


Search time

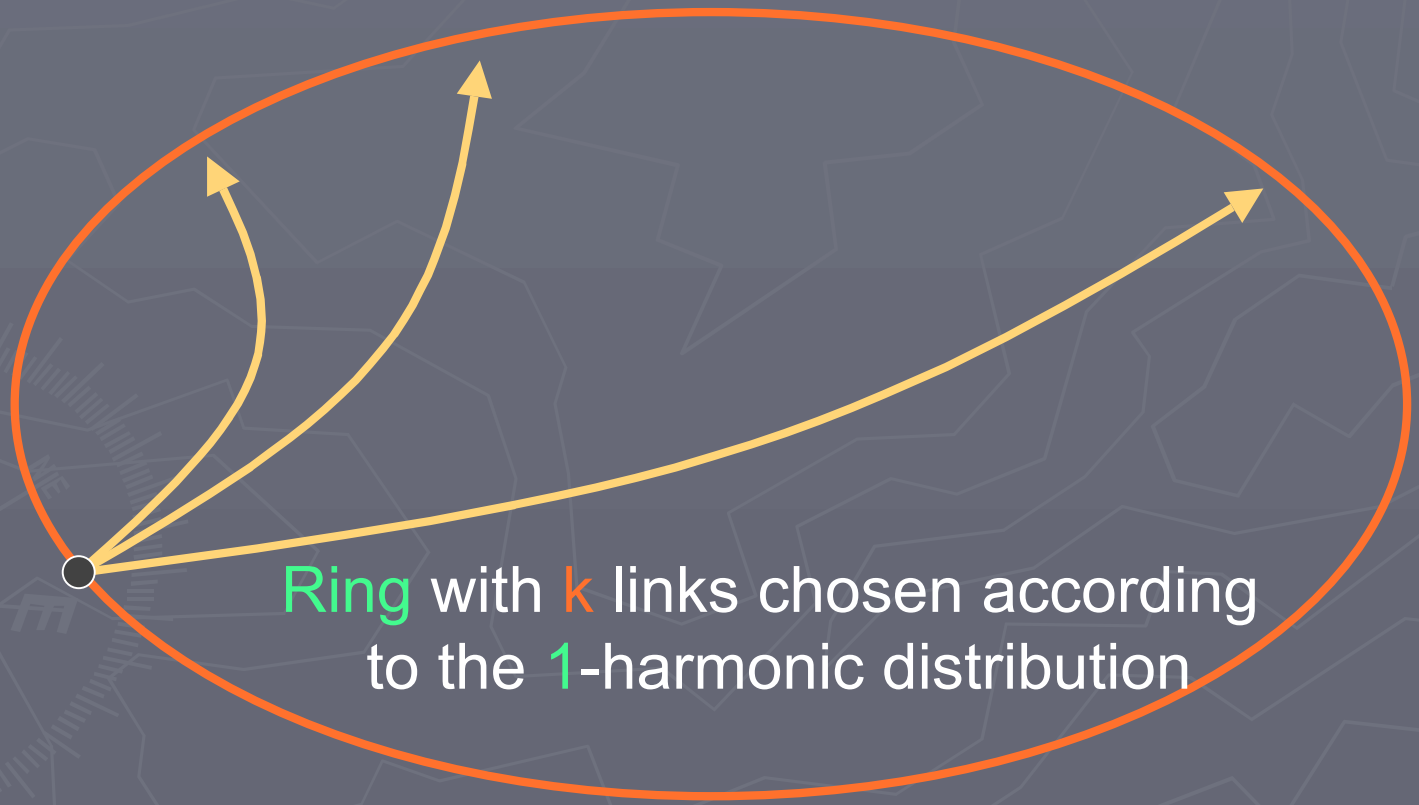

#steps
#nodes
0
50
100
150
200
250
300
350
400
450
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
50000


#steps
#nodes
0
50
100
150
200
250
300
350
400
450
100
1000
10000
100000

Search time vs. #copies



Navigable Small World DHT



Expected properties

	degree	routing
CAN	O(d)	$O(dn^{1/d})$
Chord	O(log n)	O(log n)
Viceroy	O(1)	O(log n)
D2B	O(1)	O(log n)
Small World	k	$O(\log^2 n / k)$

$1 \leq k \leq \log n$

References

References

- Go to http://www.lri.fr/~pierre
- Download
  - Peer-to-peer
    - D2B: a de Bruijn Based Content-Addressable Network
    - Combining the use of clustering and scale-free nature of user exchanges into a simple and efficient P2P system
  - Navigable networks
    - A New Perspective on the Small-World Phenomenon: Greedy Routing in Tree-Decomposed Graphs
    - Eclecticism Shrinks Even Small Worlds
- And see the references therein

THANK YOU!