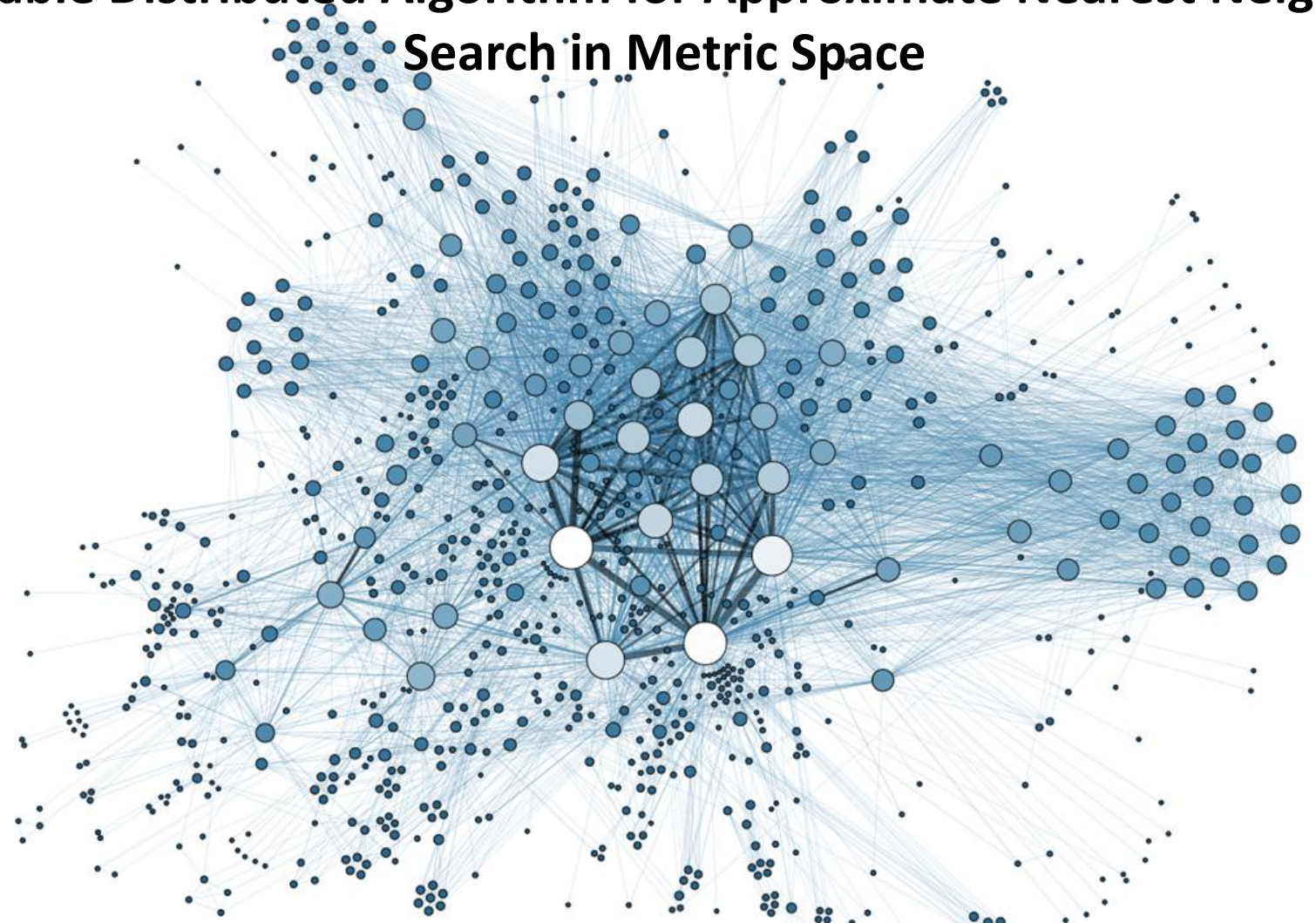
Scalable Distributed Algorithm for Approximate Nearest Neighbor Search in Metric Space



Alexander Ponomarenko, Yury Malkov, Vladimir Krylov, Logvinov Andrey
Speaker - Alexander Ponomarenko, National Research University Higher School of Economics, Laboratory of Algorithms and Technologies for Network Analysis (LATNA), Nizhny Novgorod, Russia

11 December 2014, Moscow, Russia

Motivation

- Standard approach to information retrieval on the set of complex objects is to extract the set of attributes from its and index them separately
- Indexes of the major part of information systems are based on B-trees and Hash-Tables
- This approach works well for exact search but not for the similarity search

Why is similarity?

- Any event in the history of organism is, in a sense, **unique**.
- *Recognition*, *learning*, and *judgment* presuppose an ability to categorize stimuli and classify situations by **similarity**
- Similarity (*proximity*, *resemblance*, *communality*, *representativeness*, *psychological distance*, etc.) is **fundamental** to theories of *perception*, *learning*, *judgment*, etc.

[Zezula P. et al. Similarity search: the metric space approach. – Springer, 2006. – T. 32.]

Examples with Similarity

- Does the computer disk of a suspected criminal contain illegal multimedia material?
- What are the stocks with similar price histories?
- Which companies advertise their logos in the direct TV transmission of football match?
- Is it the situation on the web getting close to any of the network attacks which resulted in significant damage in the past?

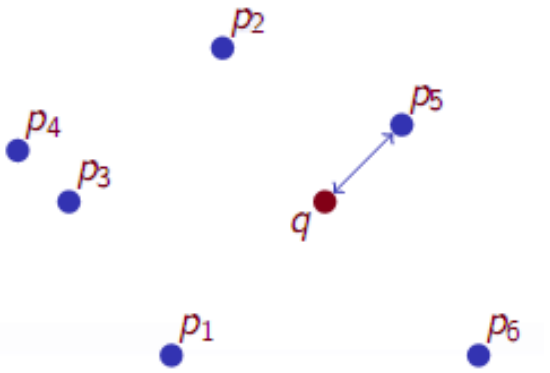
Nearest Neighbor Search

Let D – domain

$d : D \times D \to R_{[0;+\infty)}$ - distance function which satisfies properties:

- strict positiveness: d(x, y) > 0 ⇔ x ≠ y,
- symmetry: d(x, y) = d(y, x),
- reflexivity: d(x, x) = 0,
- triangle inequality: d(x, y)+d(y, z) ≥ d(x, z).

Given a finite set $X = \{p_1,\ldots,p_n\}$ of n points in some metric space (D, d), need to build a data structure on X so that for a given query point $q \in D$ one can find a point $p \in X$ which minimizes $d(p, q)$ *with as few distance computations as possible*



Applications

- pattern recognition and classification
- content-based retrieval
- machine learning
- recommendation systems
- semantic document retrieval

Results

- New model of random graphs with small world properties
- New model of navigable small world graphs
- New distributed data structure for the Nearest Neighbor problem

Examples of Distance Functions

- L_p **Minkovski distance** (for vectors)
  - L_1 – city-block distance

$$L_1(x, y) = \sum_{i=1}^{n} | x_i - y_i |$$

  - L_2 – Euclidean distance

$$L_2(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

  - L_∞ – infinity

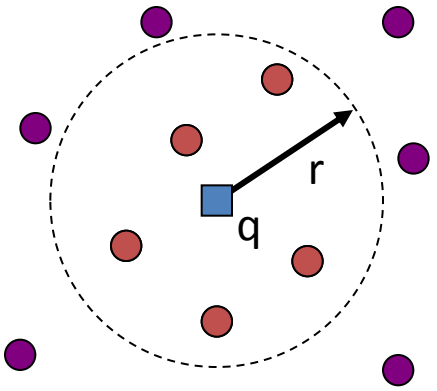
$$L_\infty(x, y) = \max_{i=1}^{n} |x_i - y_i|$$

- **Edit distance** (for strings)
  - minimal number of insertions, deletions and substitutions
  - d('application', 'applet') = 6

- **Jaccard's coefficient** (for sets A,B)

$$d(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Range Query



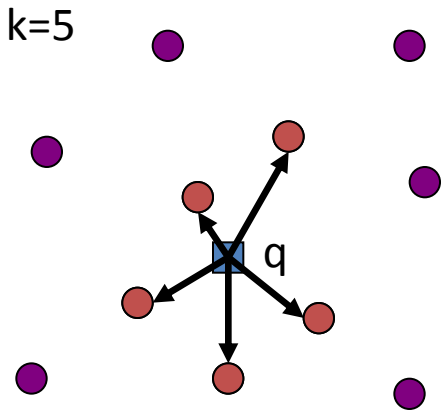
- range query
  - $R(q,r) = \{ x \in X \mid d(q,x) \leq r \}$

… all museums up to 2km from my hotel …

Nearest Neighbor Query

- the nearest neighbor query
  - $NN(q) = x$
  - $x \in X, \forall y \in X, d(q,x) \leq d(q,y)$

- k-nearest neighbor query
  - $k\text{-}NN(q,k) = A$
  - $A \subseteq X, |A| = k$
  - $\forall x \in A, y \in X - A, d(q,x) \leq d(q,y)$



… five closest museums to my hotel …

CENTRALIZED EXACT NEAREST NEIGHBOUR SEARCH STRUCTURES

SphereRectangleTree, k-d-Btree, Geometricnear-neighbor access tree Excluded, Middle vantage point forest mvp-tree, Fixed-height fixed-queries, tree Vantage-point tree, R∗-tree, Burkhard-Keller tree, BBD tree, Voronoi tree, Balanced aspect ratio tree, Metric tree, vps-tree, M-tree, SS-tree, R-tree, Spatial approximation tree, Multi-vantage point tree, Bisector tree, mb-tree, Generalized hyperplane tree, Hybrid tree, Slim tree, Spill Tree, Fixed queries tree, k-d tree, ball-tree, quadtree

The curse of dimensionality`

DISTRIBUTED EXACT NEAREST NEIGHBOUR SEARCH STRUCTURES

MAAN , SCRAP, Mercury, Voronet - only for vector space
M-Chord, GHT, MCAN – abstract metric space

Approximate Nearest Neighbor

ε-approximate:
Find a point $p \in X$ that is an ε-approximate nearest neighbor of the query q i.e. find $p \in X$ such that $d(p, q) < (1 + \varepsilon)\, d(p', q)$, where p' is the true nearest neighbor.

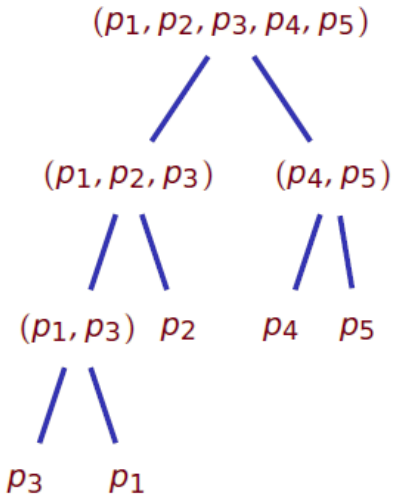
Probability approximate:
Find the true nearest neighbor with some probability ϕ.

List of methods for approximate nearest neighbor search

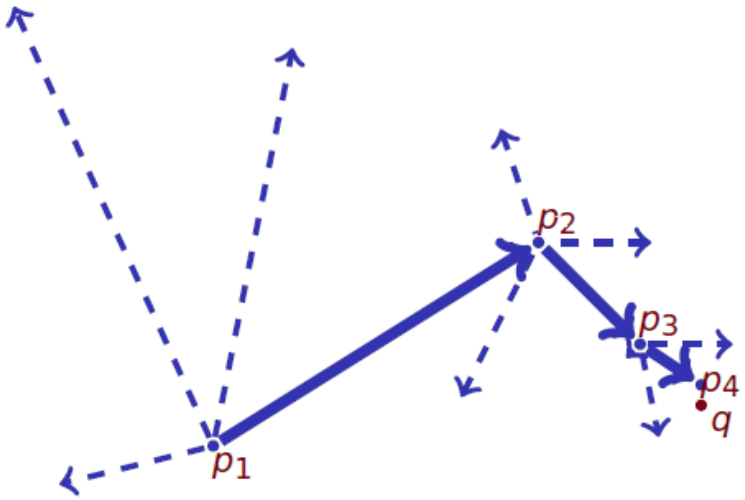
Name of the data structure	Year
Vantage Point Tree	1993
Locality Sensetive Hasing	2004
Metrized Small World	2011
Randomized KD-Tree	2008
Fast Map	1995
Metric Map	2005
M-tree: Relative Error	1998
M-tree: Good Fraction	1998
M-tree: Good Fraction	1998
M(S,Q)	1999
PAC	2000
Distinctive NN	2001
Probabilistic Proximity Search	2001
Proximity-based	2003
Probabilistic Incremental Search	2004
Approximate k-NN with Antipole Tree	2005
Approximate/On-Line NN: Distance Ratio	2001
Genetic Search	2007
Anytime k-NN Search	2008
Arwalk algorithm	2008
Permutation Index	2008
Permutation Index with incremental sorting	2008
Permutation prefix tree	2012
Permustaion with inverted index	2008
Permutation with metric index (vp-tree)	2009

Three Famous Technique

Branch and bound

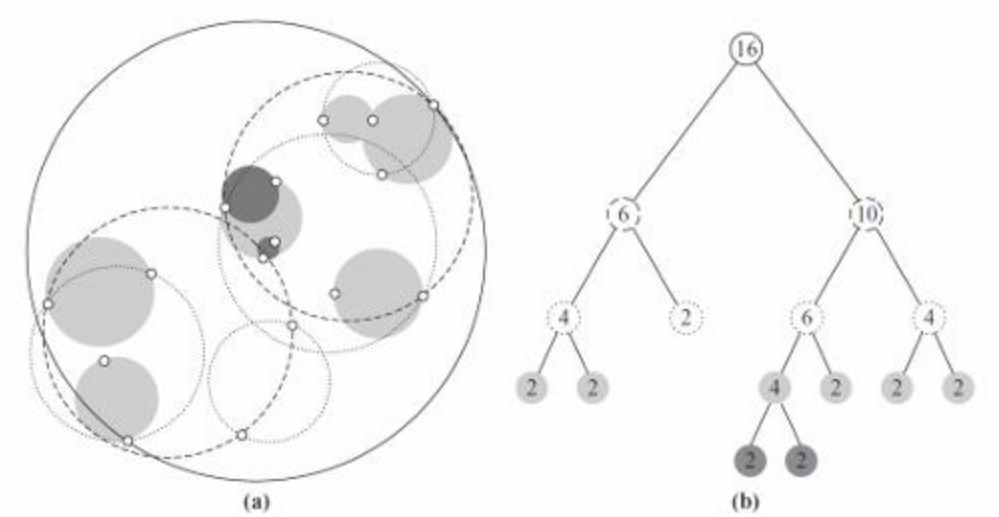


Greedy walks



Mappings: LSH,
random projections, minhashing

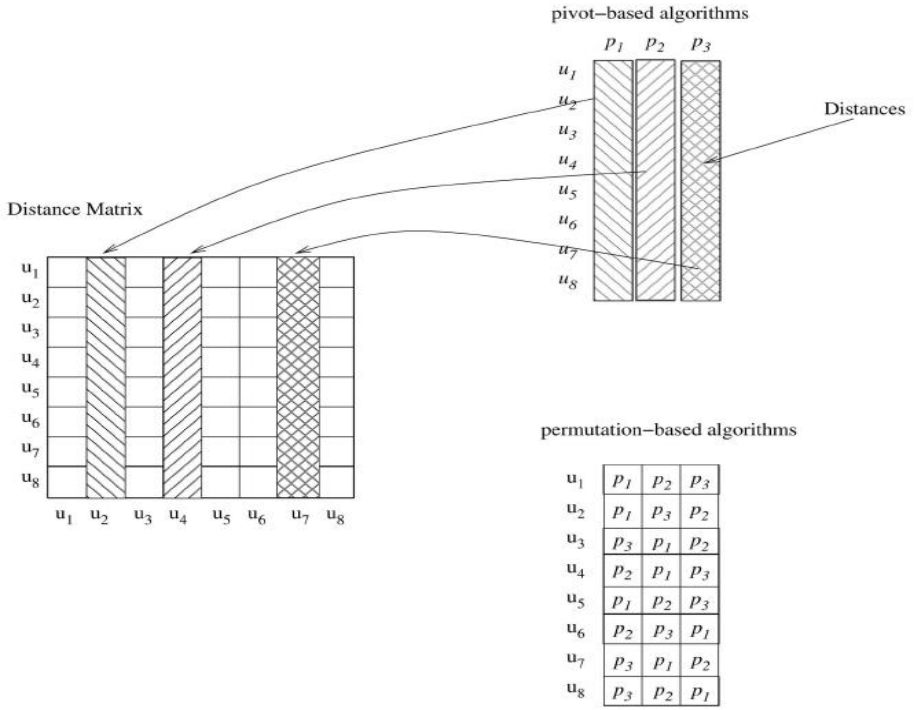
Vantage Point Tree



The Vantage Point Tree is a hierarchical space partitioning method which uses triangle inequality to discard partitions that cannot contain nearest neighbors

[Yianilos Peter N. "Data structures and algorithms for nearest neighbor search in general metric spaces." Proceedings of the fourth annual ACM-SIAM Symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 1993]

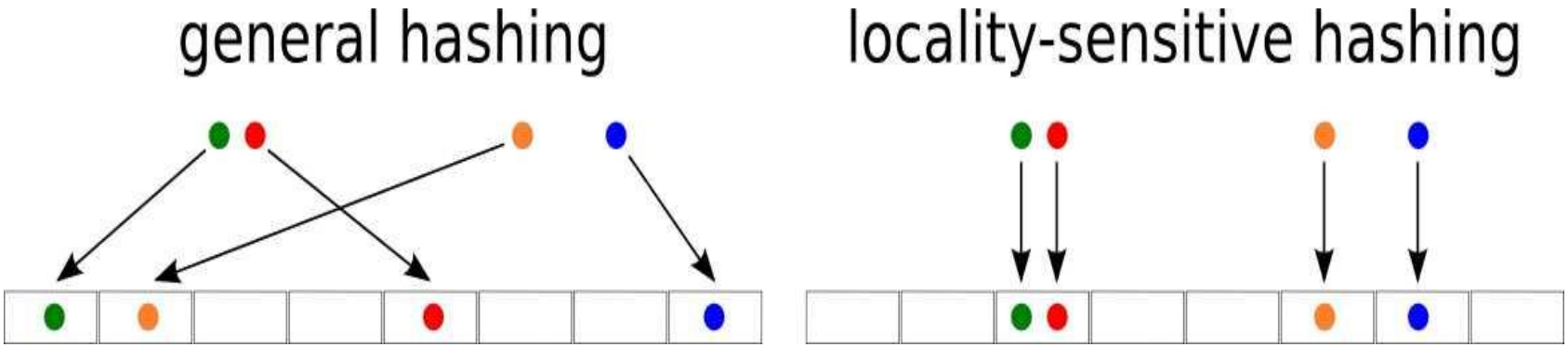
Permutation Index



Permutation methods are dimensionality-reduction approaches, where each point is represented by a low-dimensional integer-valued vector called a *permutation*. A permutation Π_x of $(1...k)$ is an ordered set of pivots $\{p_1, p_2, ..., p_k\}$, where the order depends on the distance between pivot and object i.e. $d(p_{\Pi_x(i)}, x) < d(p_{\Pi_x(i+1)}, x)$

[Gonzalez Edgar Chavez, Karina Figueroa and Gonzalo Navarro. "Effective proximity retrieval by ordering permutations." Pattern Analysis and Machine Intelligence, IEEE Transactions on 30.9 (2008): 1647-1658]

Locality Sensitive Hashing



The general idea of hashing is to avoid collisions. The idea of LSH is to exploit collisions for mapping points which are nearby (in geometrical sense) into the same bucket.

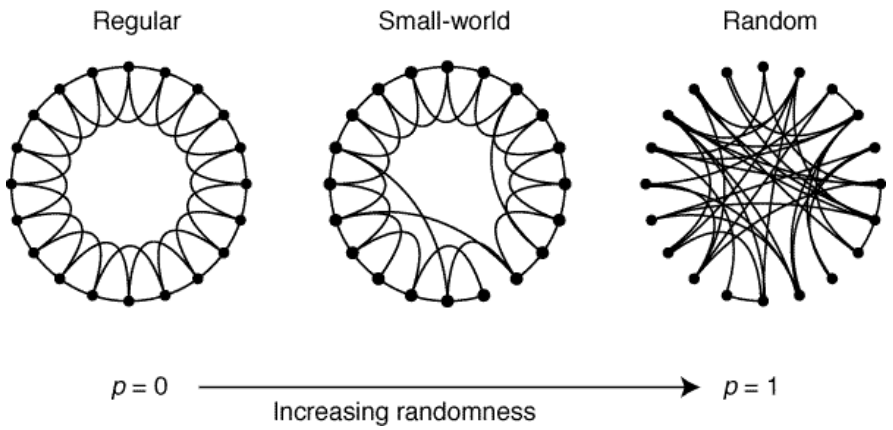
[P. Indyk and R. Motwani, “Approximate nearest neighbors: towards removing the curse of dimensionality,” in Proceedings of the thirtieth annual ACM symposium on Theory of computing. ACM, 1998, pp. 604–613]

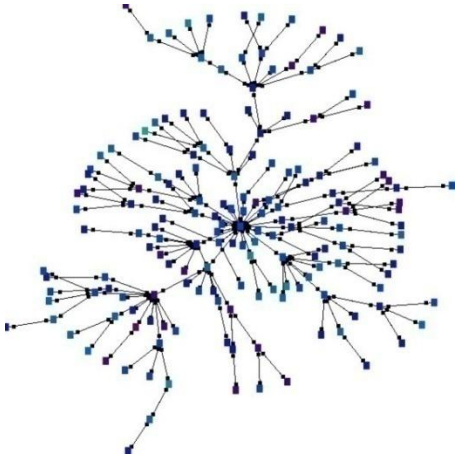
Our goal to construct the network such that:

- every node knows only small number of other nodes
- search can be started from any node
- search uses only local information
- the expected number of steps to rich destination is polylogarithmic from the number of all nodes

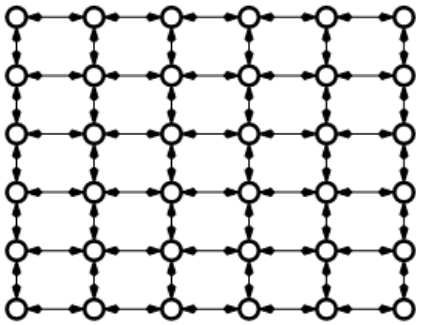
The Small World Networks

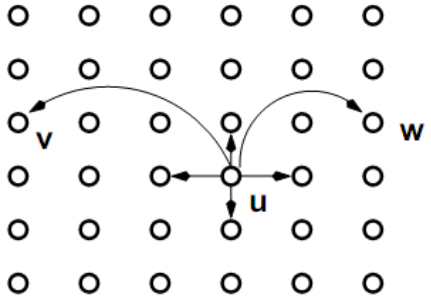
Two famous “Blind” models: “Watts-Strogatz” and “Barabási–Albert”



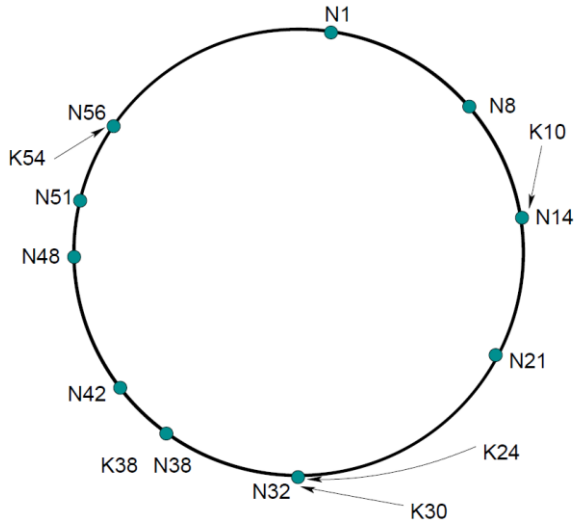


Navigable small world model of Kleinberg





Structured Peer-to-Peer Networks: Chord Protocol

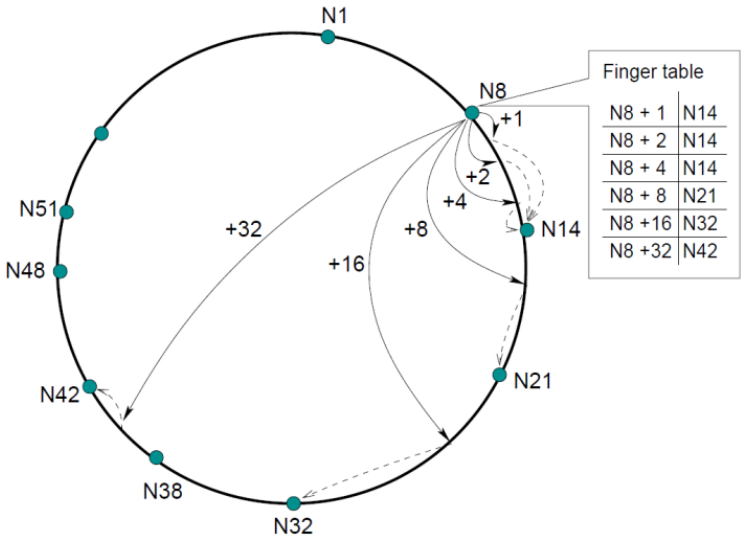


An identifier circle consisting of 10 nodes storing five keys.

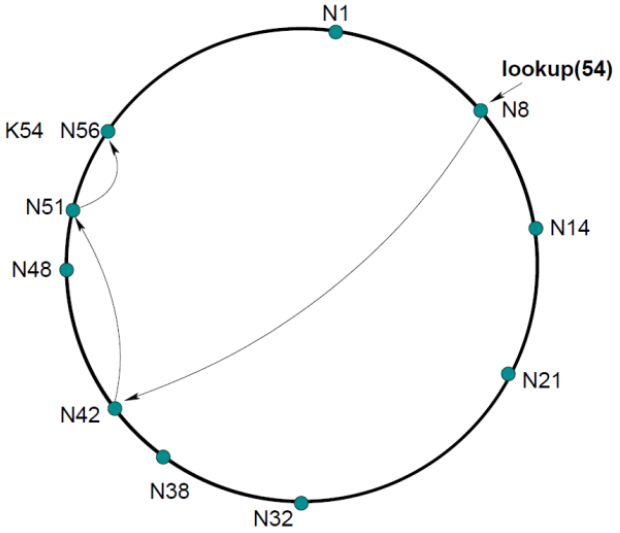
Identifiers are ordered in an identifier circle modulo 2^m.
Key k is assigned to the first node whose identifier is equal to or follows k in the identifier space. This node is called the successor node of key k, denoted by successor(k). If identifiers are represented as a circle of numbers from 0 to 2^m - 1, then successor(k) is the first node clockwise from k.

[Stoica I. et al. Chord: a scalable peer-to-peer lookup protocol for internet applications //Networking, IEEE/ACM Transactions on. – 2003. – Т. 11. – №. 1. – С. 17-32.]

Structured Peer-to-Peer Networks: Chord Protocol



Routing table of node «N8»

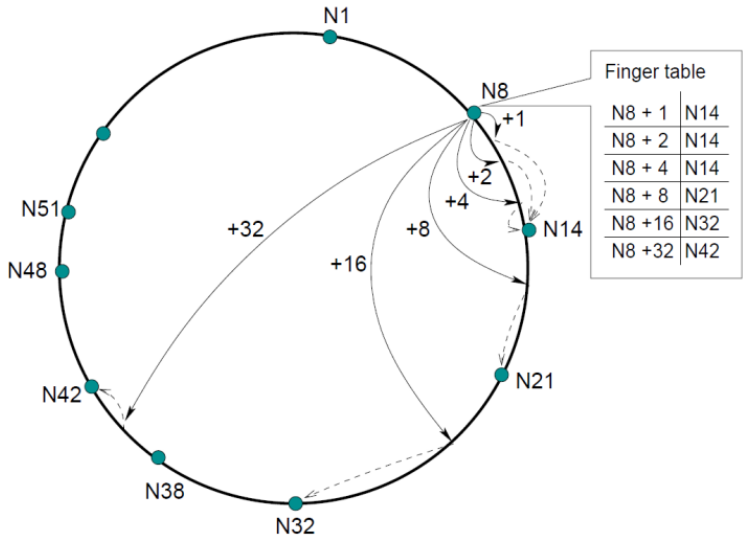


Searching of key 54 staring from «N8».

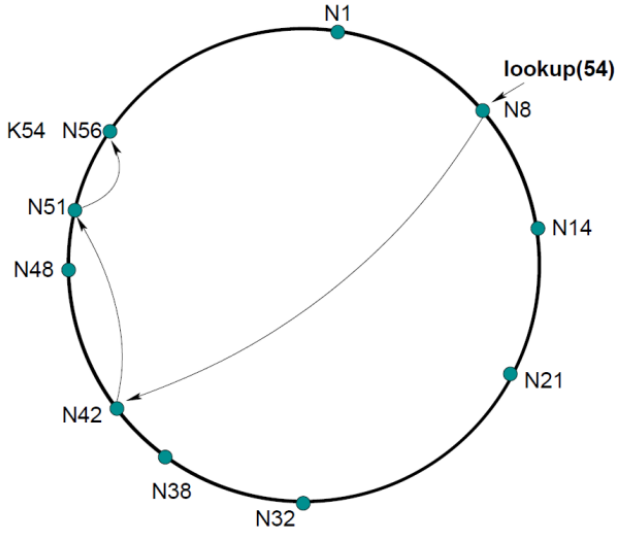
Distance function: $d(x,y)= (y - x) \mod 2^m$

Each node, *n*, maintains a routing table with (at most) *m* entries, called the *finger table*. The i-th entry in the table at node *n* contains the identity of the first node, s, that succeeds *n* by at least 2^(*i*-1) on the identifier circle, i.e., s = successor(*n*+2^(*i*-1)), where 1<= *i*<= *m*

Structured Peer-to-Peer Networks: Chord Protocol



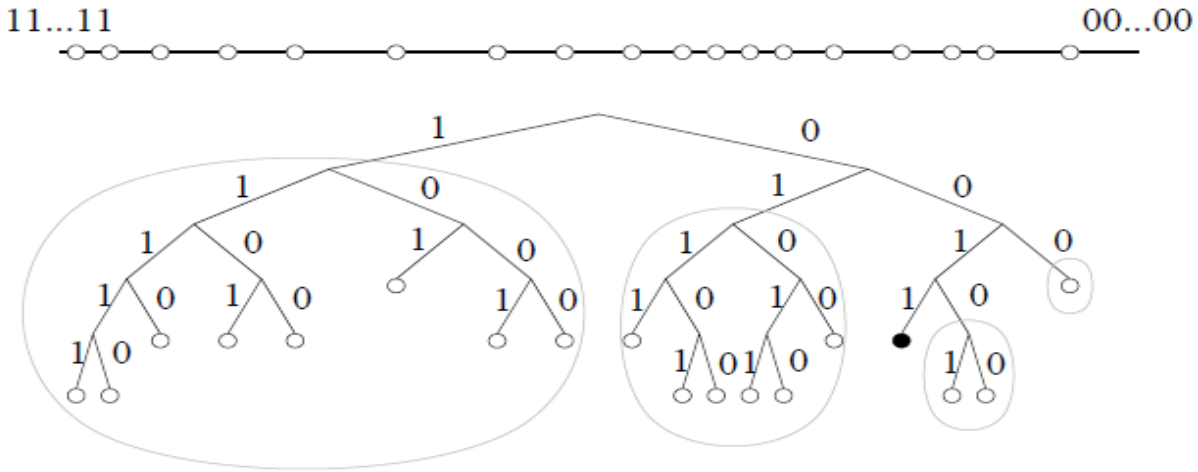
Routing table of node «N8»



Searching of key 54 staring from «N8».

Theorem: *With high probability (or under standard hardness assumptions), the number of nodes that must be contacted to find a successor in an N-node network is O(logN).*

Structured Peer-to-Peer Networks: Kademlia

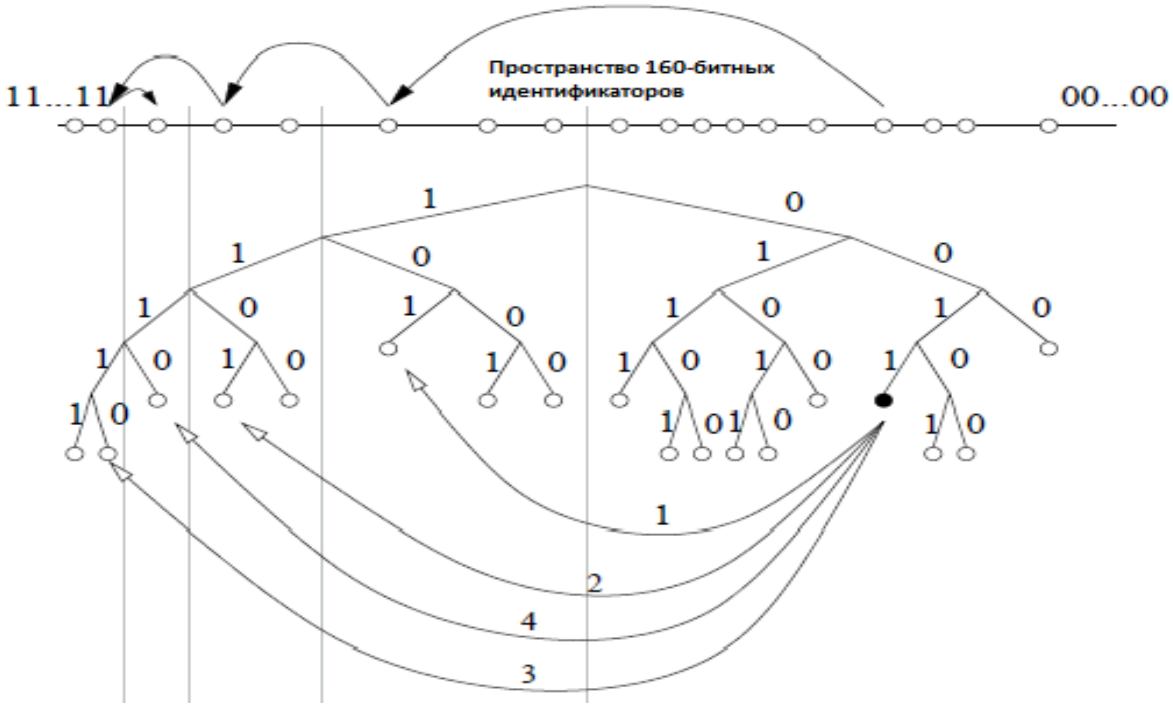


Identifier space of Kademlia

Distance function: *d(x,y)= x xor y*

Maymounkov P., Mazieres D. Kademlia: A peer-to-peer information system based on the xor metric //Peer-to-Peer Systems. – Springer Berlin Heidelberg, 2002. – С. 53-65.

Structured Peer-to-Peer Networks: Kademlia

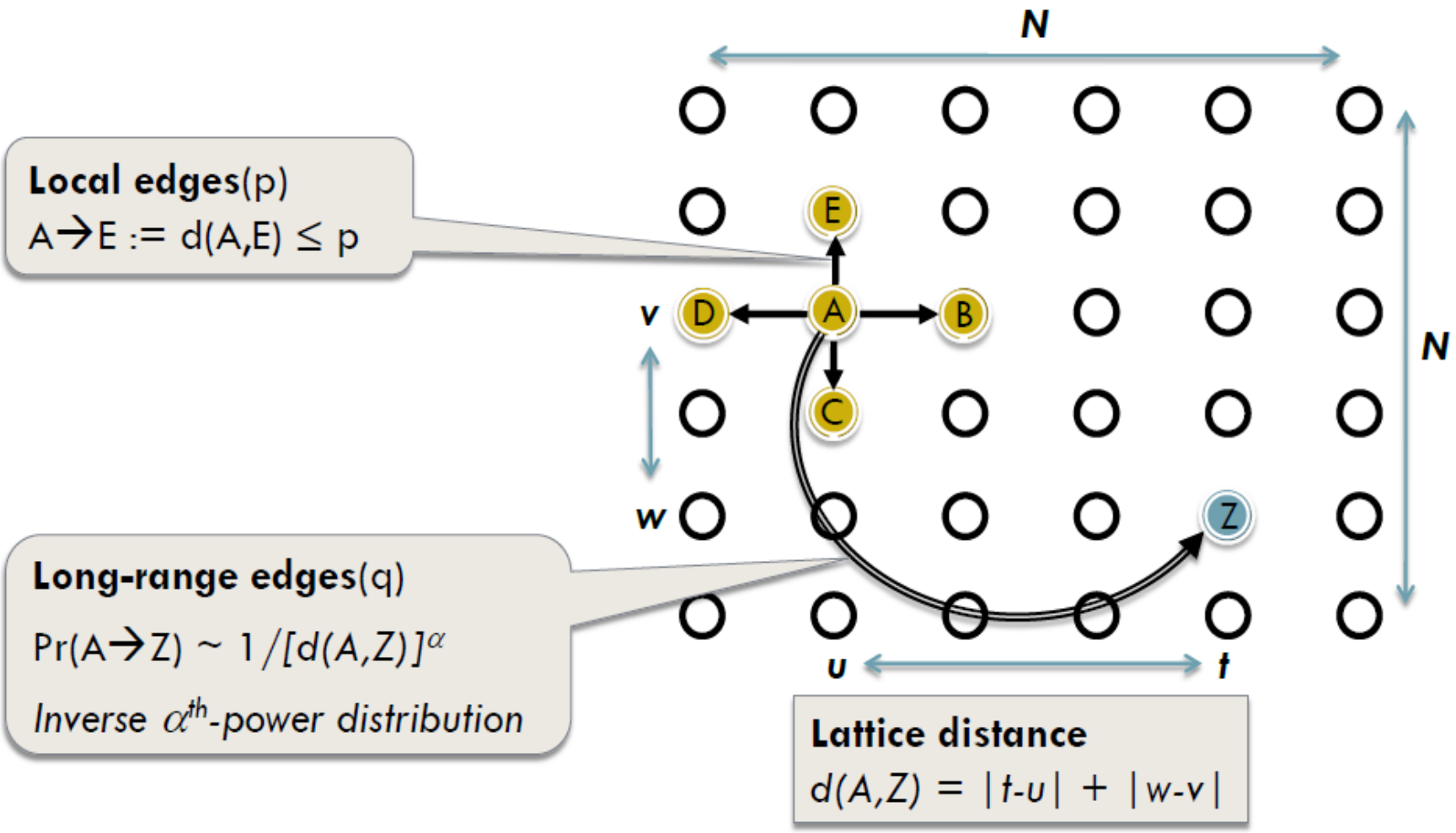


The node with unique prefix id 0011 finding node with id 1110

Structured Peer-to-Peer Networks: Applications

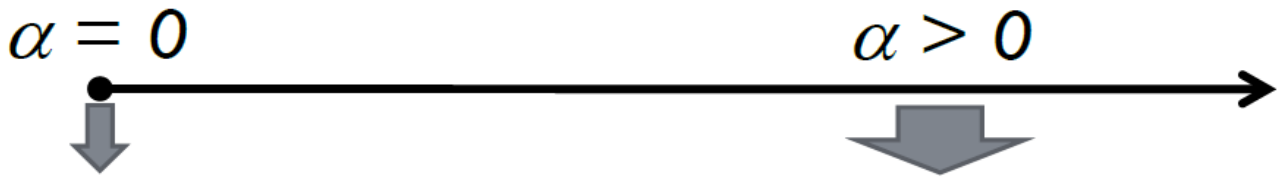
- Peer-to-Peer file sharing systems
- Key-Value Storages
- Load Balancers
- Cooperative Mirroring

Kleinberg's Navigable Small World



[Kleinberg J. The small-world phenomenon: An algorithmic perspective //Proceedings of the thirty-second annual ACM symposium on Theory of computing. – ACM, 2000. – C. 163-170.]

Family of network models with parameter α



Long-range contacts chosen independently of their position (~Watts-Strogatz model)

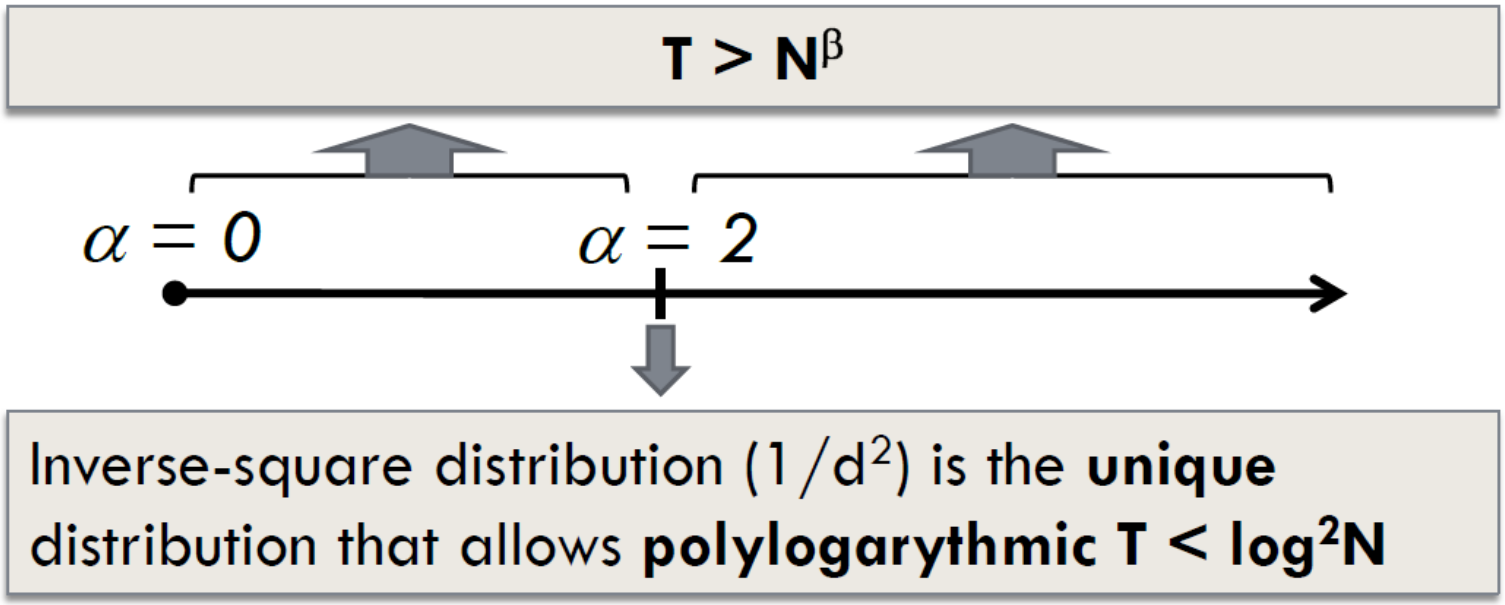
Long-range contacts tend to cluster in the nodes' vecinity

Which α *yields an* ***effectively*** *navigable network?*

Expected delivery time T

- Expected number of steps to reach the destination
- Shortness (small T) of paths is defined as ***polylogarithmic***

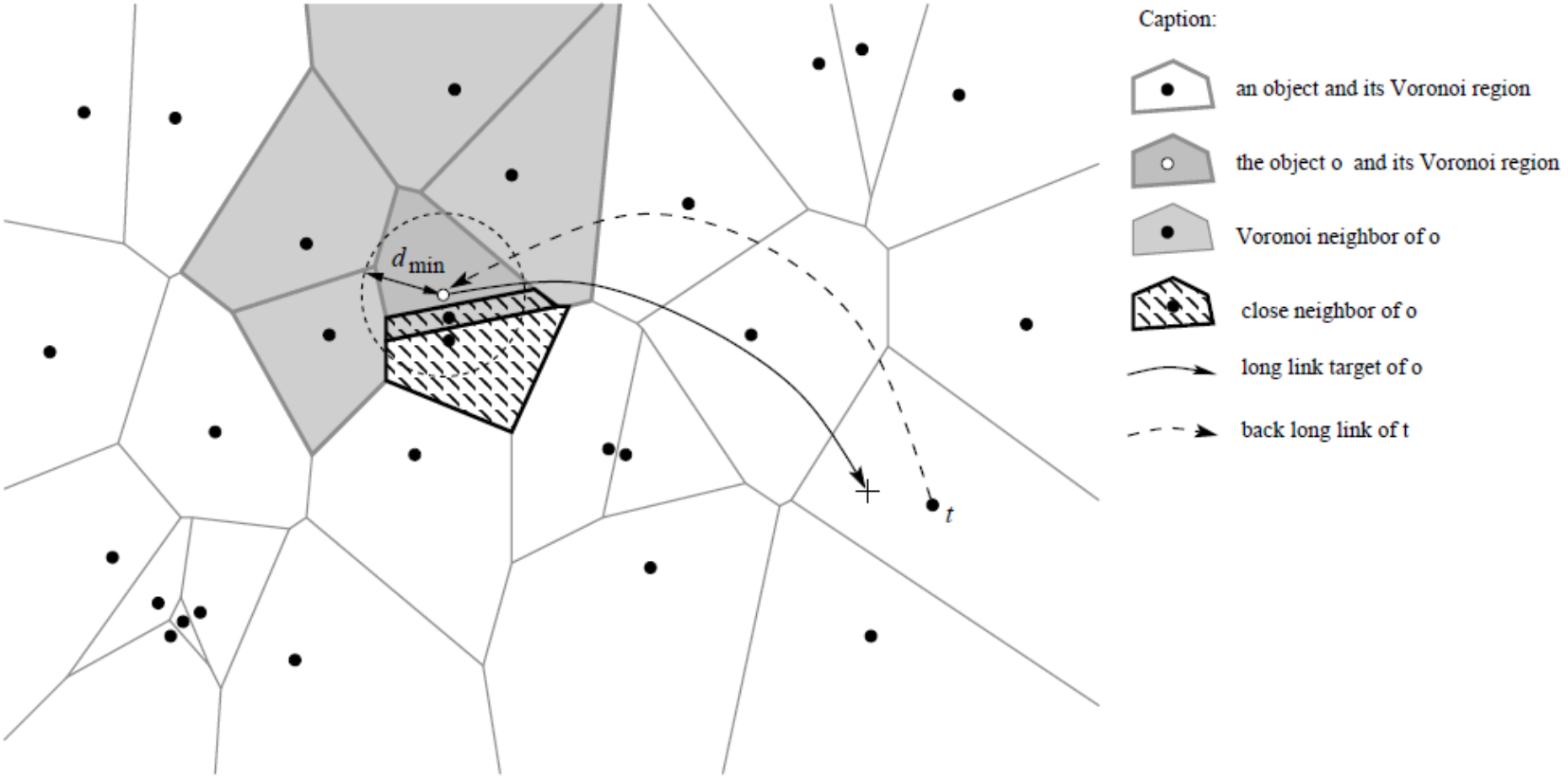
Kleinberg's Navigable Small World



Generalization

For a k-dimensional lattice, paths are polylogarithmic iff $\alpha = k$

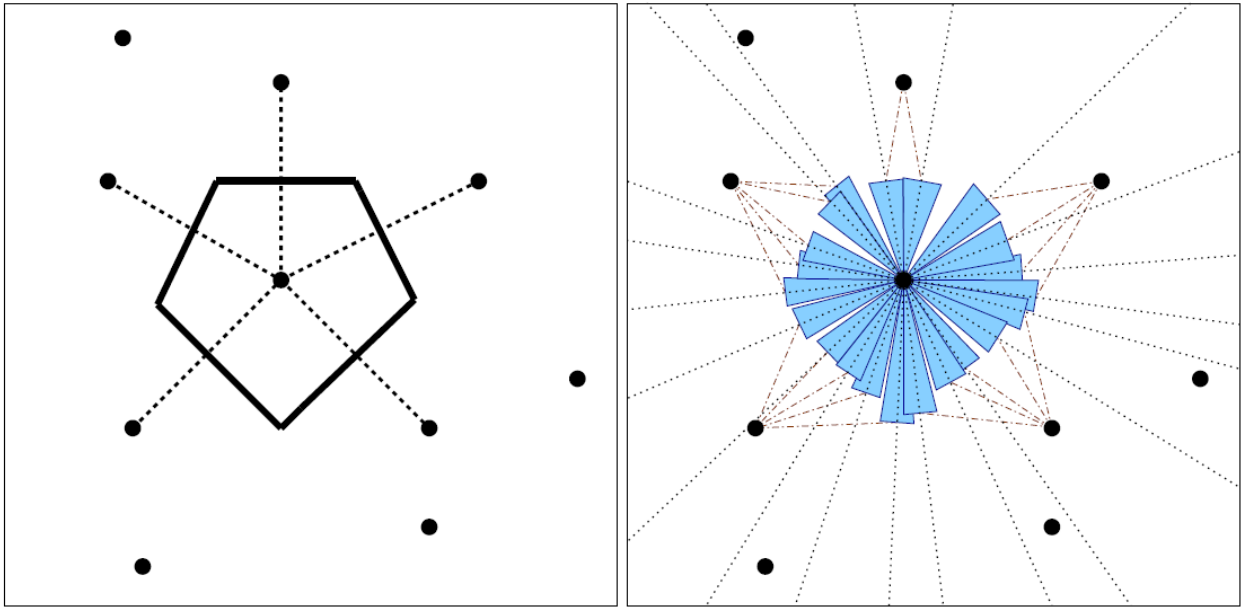
VoroNet: A scalable object network based on Voronoi tessellations



Distance function: $d(x, y) = \sqrt[2]{x^2 + y^2}$

Beaumont O. et al. VoroNet: A scalable object network based on Voronoi tessellations. – 2006.

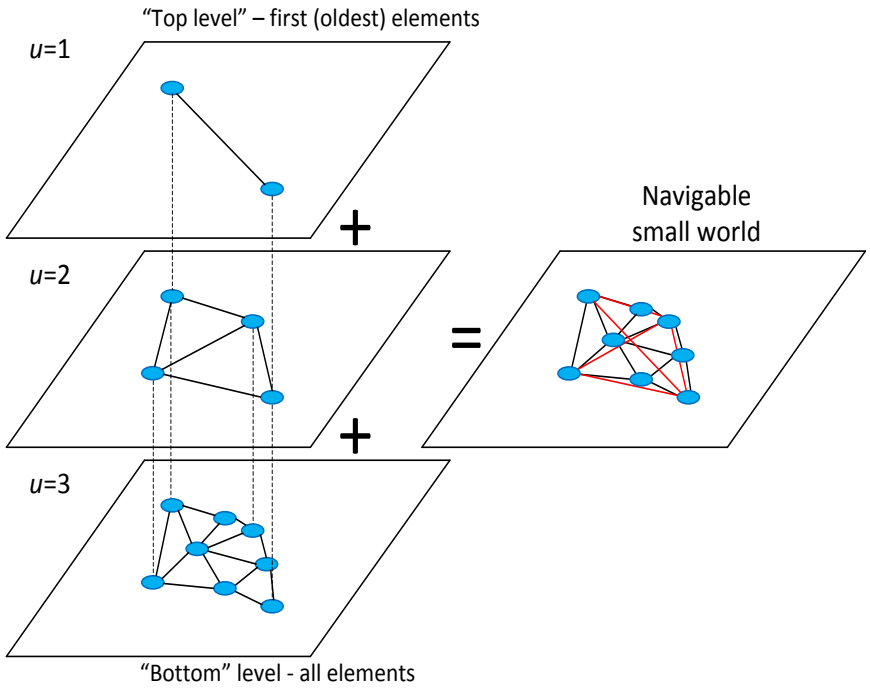
RayNet

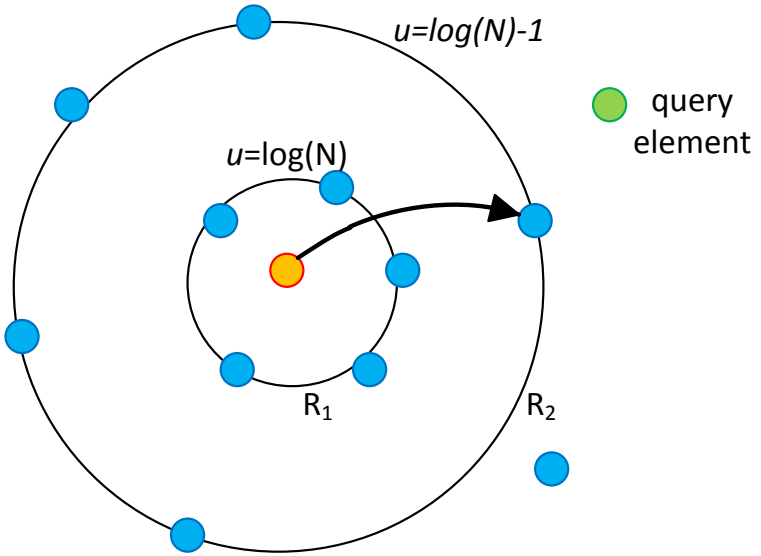


Distance function: $d(x, y) = \sqrt[D]{x^D + y^D}$

Beaumont O., Kermarrec A. M., Rivière É. Peer to peer multidimensional overlays: Approximating complex structures //Principles of Distributed Systems. – Springer Berlin Heidelberg, 2007. – С. 315-328.

Metrized Small World Algorithm





[Y. Malkov, A. Ponomarenko, A. Logvinov, and V. Krylov, "Scalable distributed algorithm for approximate nearest neighbor search problem in high dimensional general metric spaces," in Similarity Search and Applications. Springer, 2012, pp. 32–147]

[Y. Malkov, A. Ponomarenko, A. Logvinov, and V. Krylov, "Approximate nearest neighbor algorithm based on navigable small world graphs," Information Systems, vol. 45, 2014, pp. 61–68.]

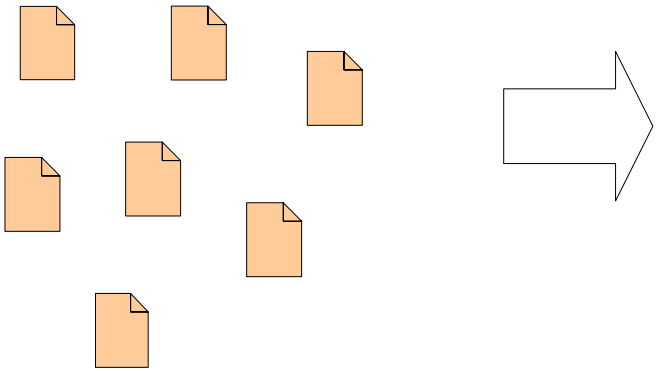
Structure overview

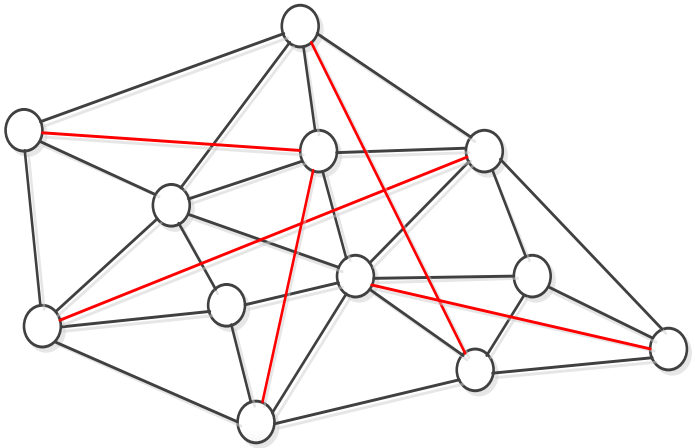
Search algorithm: greedy walk + multisearch
Structure: graph with navigable small word topology

Two level links:

Short links -> correct navigation

Long links -> fast navigation

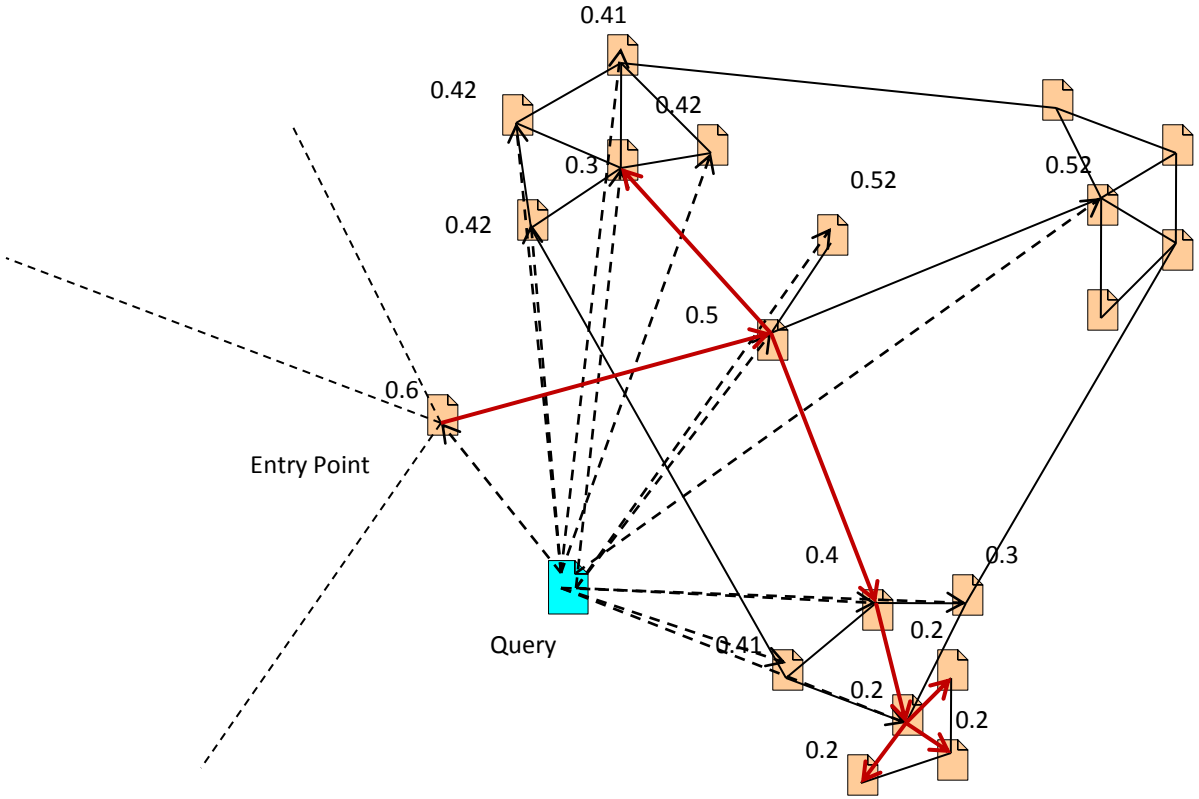




$X = \{p_1,\ldots,p_n\}$

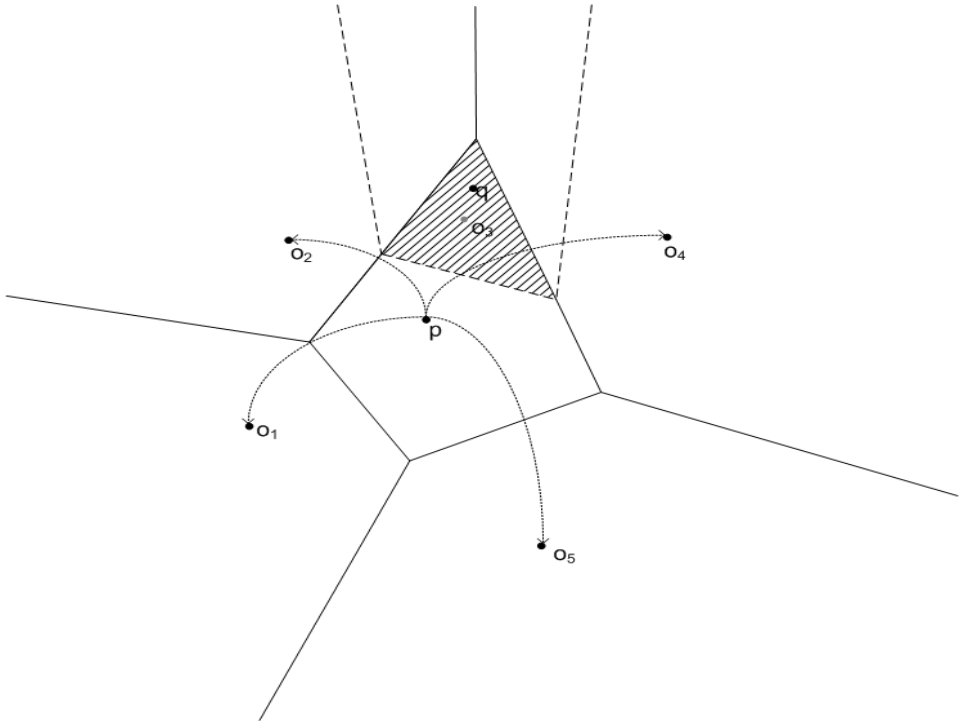
G(V,E)

Search by greedy algorithm



How can we make this greedy search to work correct?

-Graph of the network should contains the Delaunay subgraph



$$\forall a \in V, \forall q \in U, \text{if } \forall b \in N(a), f(q,a) \leq f(q,b), \text{ then } \forall b \in V, f(q,a) \leq f(q,b).$$

Curse of Dimensionality

The number of Voronoi neighbors has exponential rising with increasing number of dimensionality.

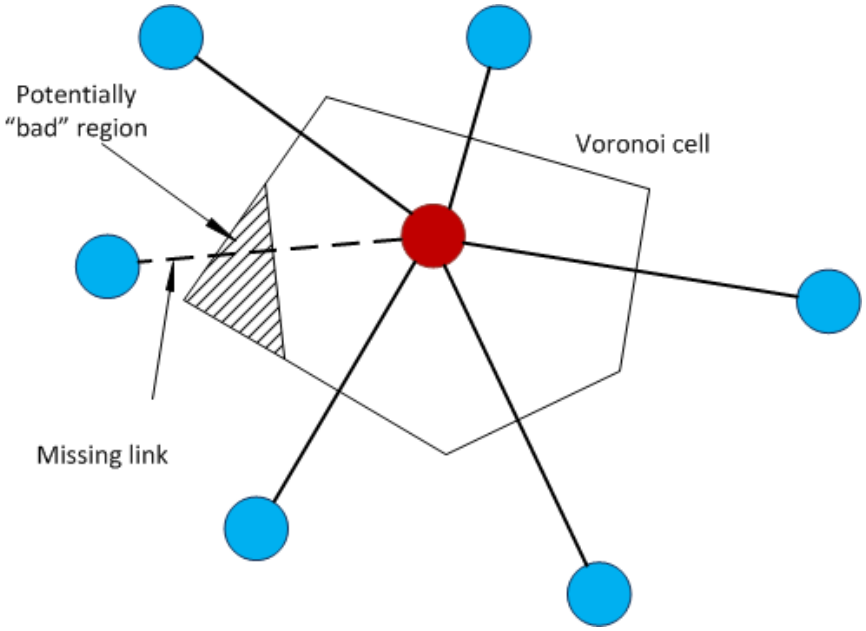
That is impossible find all Voronoi neighbors in arbitrary Metric Space using only metric calculation. [Gonzalo Navaro 1999 “Searching in Metric Spaces by Spatial Approximation”]

Theorem: *given a set S of elements in an unknown metric space U, and given the distances among each pair of elements in S, then for each $a, b \in S$ there exists a valid metric space U where a and b are connected in the Delaunay graph of S.*

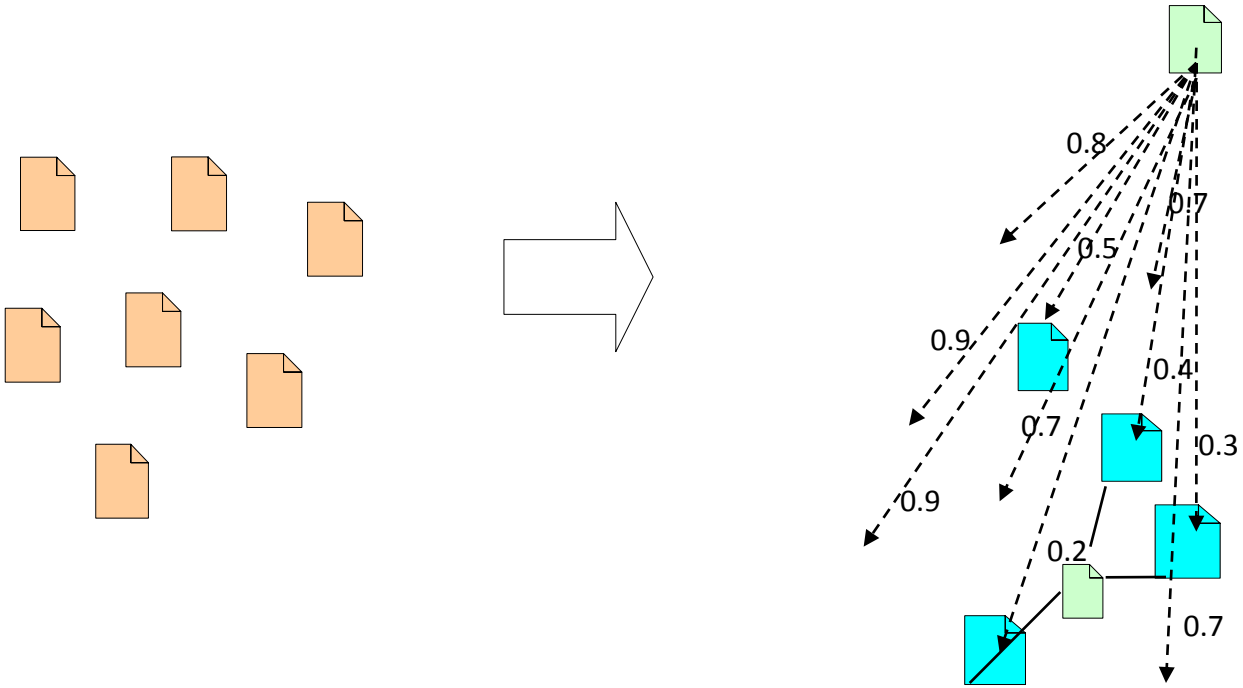
Building the short links

The main goal is to minimize the probability of the false local minima while the keeping number of links small.

We propose to assemble the structure by adding elements one by one and connecting them on each step with the K closest objects which are already in the structure.



Construction algorithm



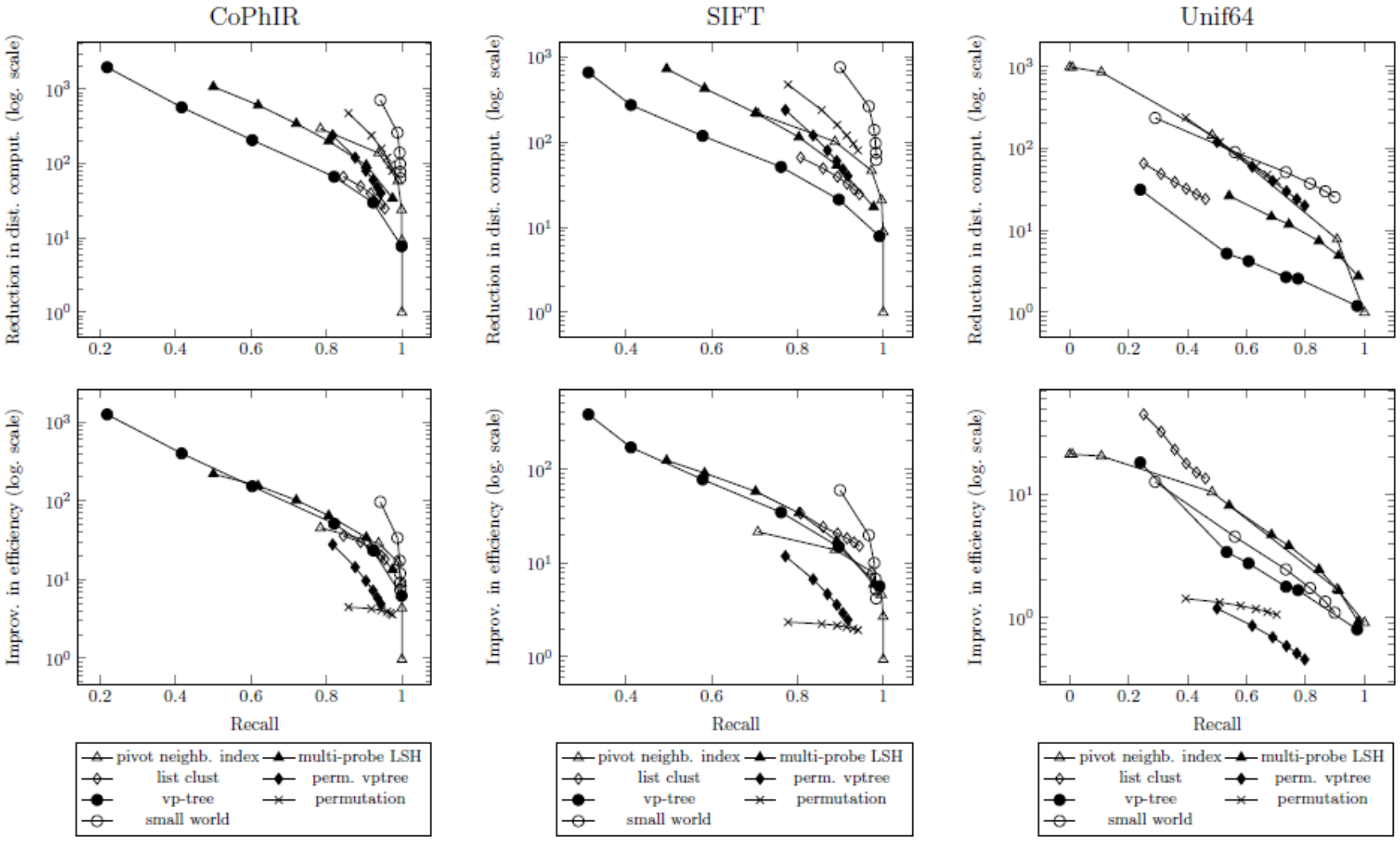
Data sets

- CoPhIR (L2) is the collection of 208-dimensional vectors extracted from images in MPEG7 format.
- SIFT is a part of the TexMex dataset collection available http://corpus-texmex.irisa.fr It has one million 128-dimensional vectors. Each vector corresponds to descriptor extracted from image data using Scale Invariant Feature Transformation (SIFT)
- Unfi64 is synthetic dataset of 64-dimensional vectors. The vectors were generated randomly, independently and uniformly in the unit hypercube.

Recall and precision measures

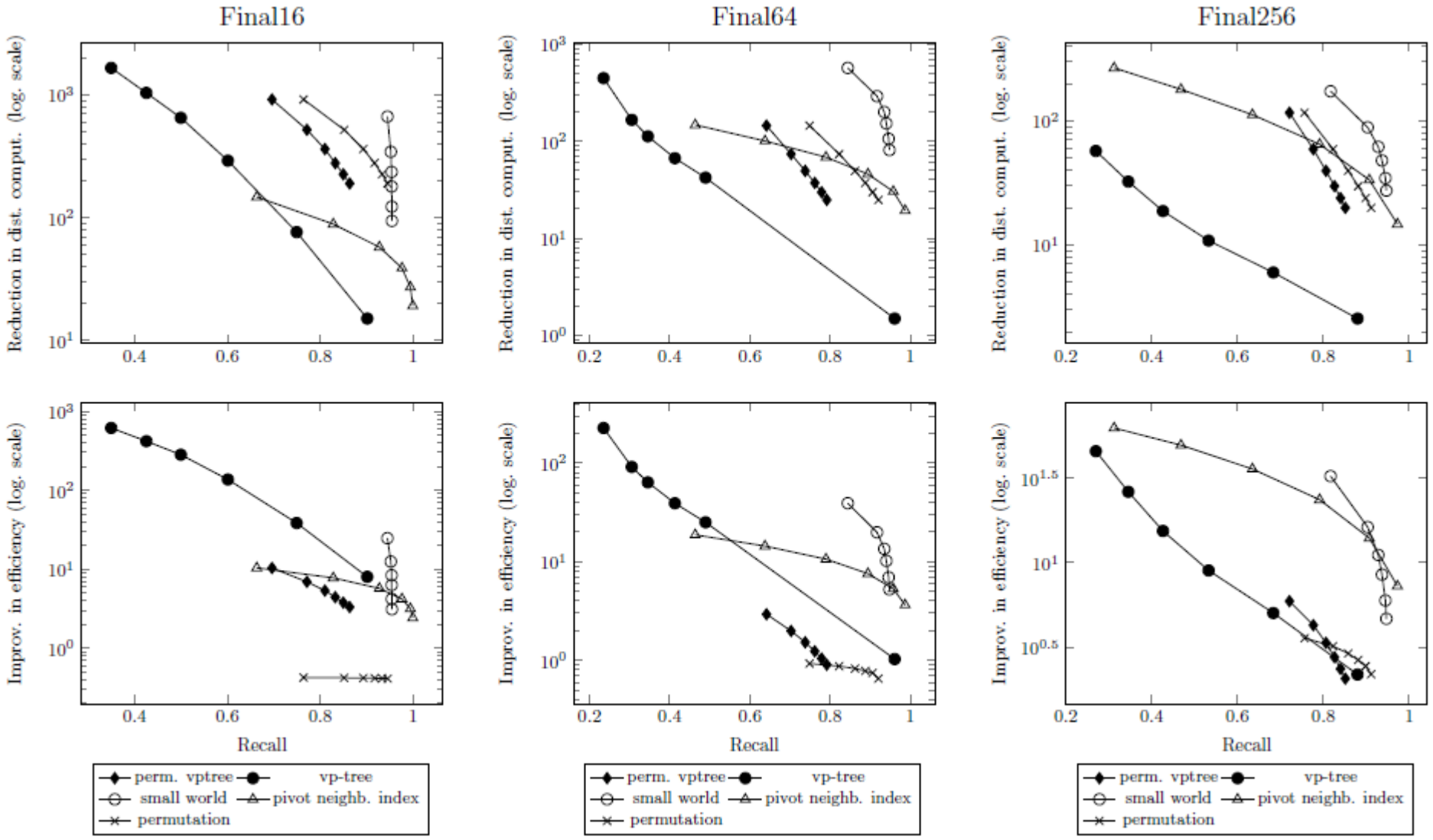
$$recall = \frac{relevant\_retrieved}{all\_relevant}$$

$$precision = \frac{relevant\_retrieved}{all\_retrieved}$$



Performance of a 10-NN search for L_2: plots in the same column correspond to the same data set

[Ponomarenko A. et al. Comparative Analysis of Data Structures for Approximate Nearest Neighbor Search //DATA ANALYTICS 2014, The Third International Conference on Data Analytics. – 2014. – C. 125-130.]



KL-divergence: $d(x, y) = \sum x_i \log \frac{x_i}{y_i}$

Final16, Final64, and Final256: are sets of 0.5 million topic histograms generated using the Latent Dirichlet Allocation (LDA).

Wikipedia dataset

Vector Space Model

$$d_j = (w_{1,j}, w_{2,j}, ..., w_{n,j})$$

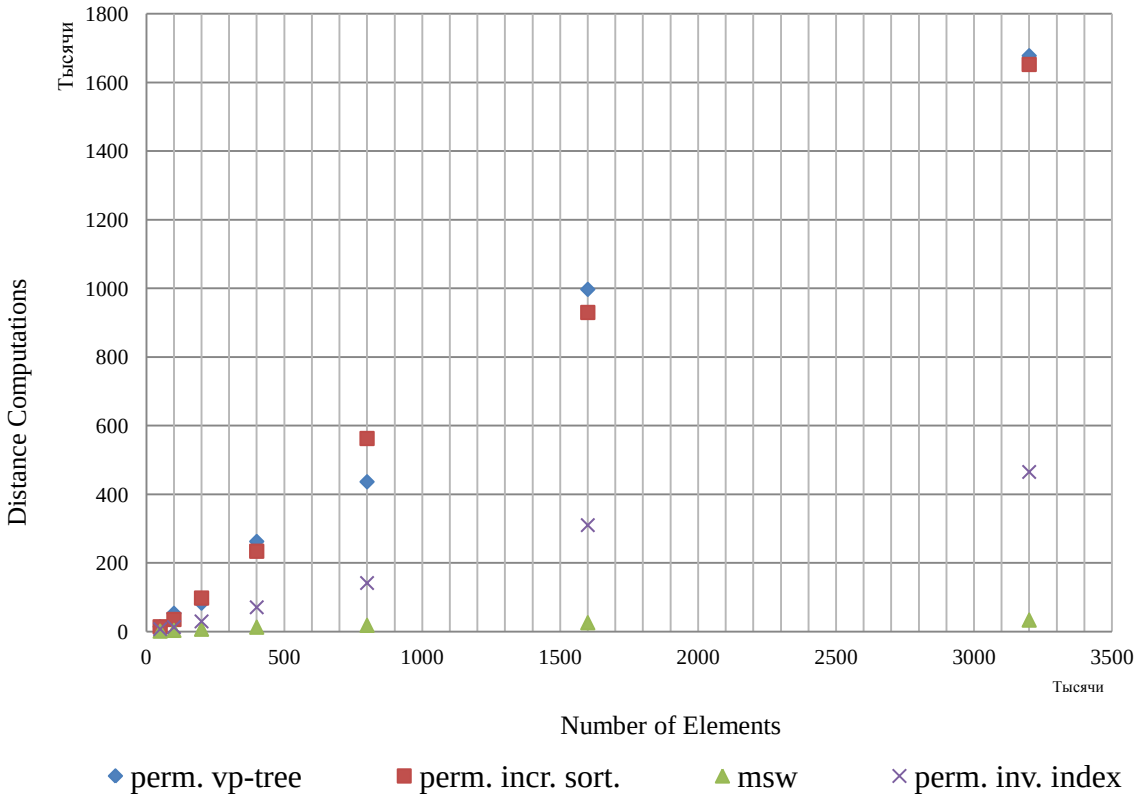
$$q = (w_{1,q}, w_{2,q}, ..., w_{n,q})$$

$$sim(d_j, q) = \frac{d_j \cdot q}{\| d_j \| \cdot \| q \|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \sqrt{\sum_{i=1}^{n} w_{i,q}^2}}$$

Wikipedia (cosine similarity): is a data set that contains 3.2 million vectors represented in a sparse format.
This set has an extremely high dimensionality (more than 100 thousand elements). Yet, the vectors are sparse: On average only about 600 elements are non-zero.

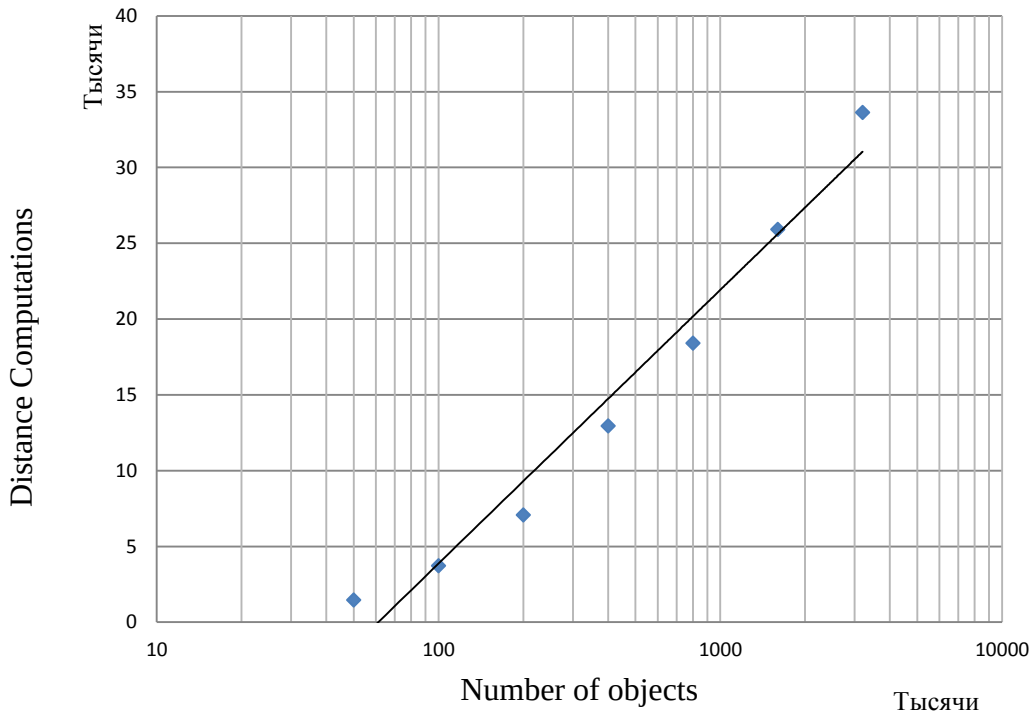
Scaling of methods on Wikipedia dataset

Recall = 0.9

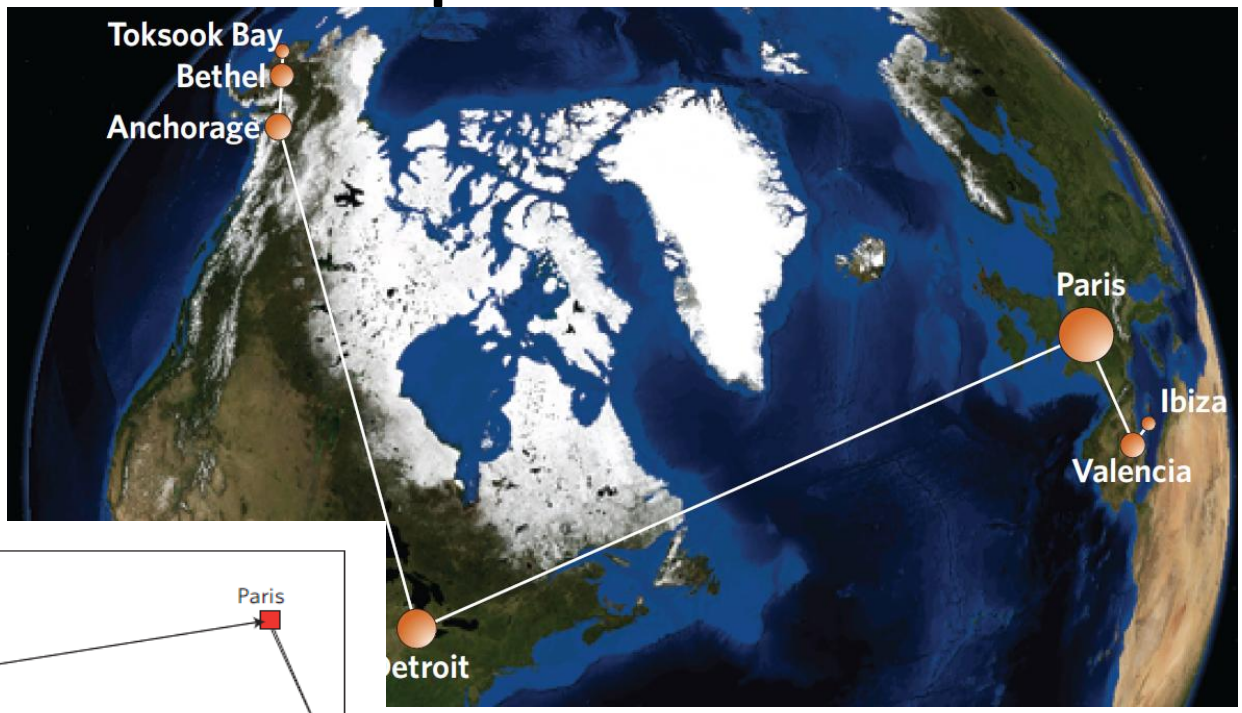


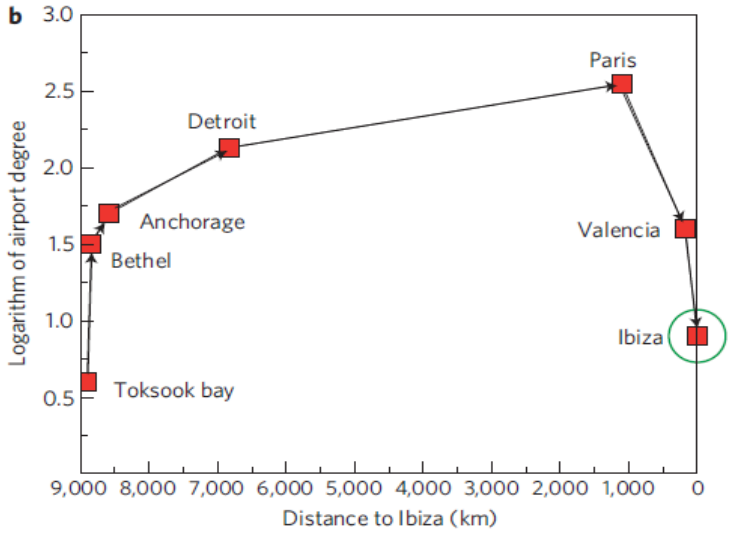
Wikipedia is dataset that contains 3.2 million vectors represented in a sparse format. Each vector corresponds to the frequency term vector of the Wikipedia page extracted using the gensim library. This set has an extremely high dimensionality (more than 100 thousand elements).

Scaling of MSW data structure



Air travel by greedy routing as an explanation





[Boguna M., Krioukov D., Claffy K. C. Navigability of complex networks //Nature Physics. – 2008. – Т. 5. – №. 1. – С. 74-80.]

Summing up

- Algorithm is very simple
- Algorithm uses only distance values between the objects, making it suitable for arbitrary spaces.
- Proposed data structure has no root element.
- All operations (addition and search) use only local information and can be initiated from any element that was previously added to the structure.
- Accuracy of the approximate search can be tuned without rebuilding data structure
- Algorithm high scalable both in size and data dimensionality

Good base for building many real-world extreme dataset size high dimensionality similarity search applications