INTELLIGENT INFORMATION RETRIEVAL IN DISTRIBUTED SYSTEMS

Prof. Dr. Vesna Šešum-Čavić

Faculty of Civil Engineering

University of Belgrade

vsesumcavic@grf.bg.ac.rs

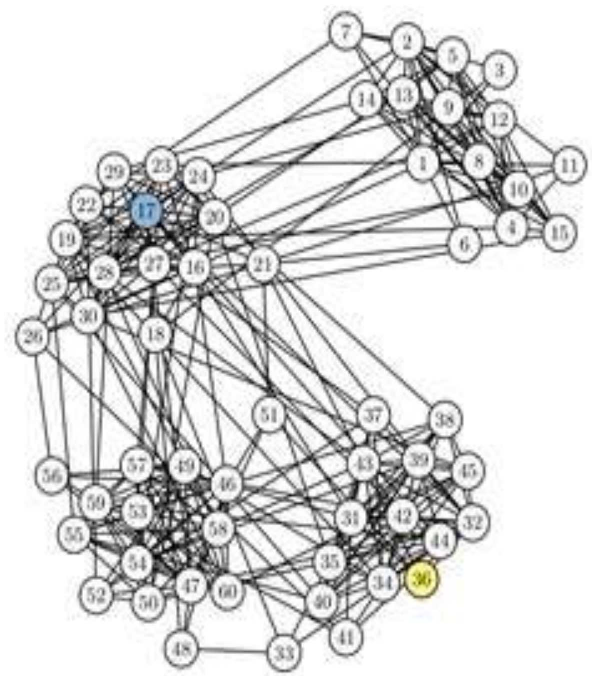
www.complang.tuwien.ac.at/vesna

Overview

- Motivation: Complexity in Distributed Systems
- Benefits of Swarm-Intelligence Approach
- Case Study: Location and Retrieval of Information
- Summary

Complexity in Distributed Systems

- Forced to integrate other software systems and components that are often not reliable, exhibit bad performance, and are sometimes unavailable [6]
- How to cope with huge dynamics and vast number of unpredictable dependencies on participating components?
- Autonomously acting components inspired by nature
- The unavoidable complexity cannot be eliminated
  - the complexity can be shifted



Complexity in Distributed Systems [4]

Swarm Intelligence (1)

- A self-organizing biological system [6]
- A swarm can be defined as a structured collection of interacting organisms (or agents)
- The collective behaviors of (unsophisticated) agents interacting locally with their environment cause coherent functional global patterns to emerge
- Distributive and autonomous properties

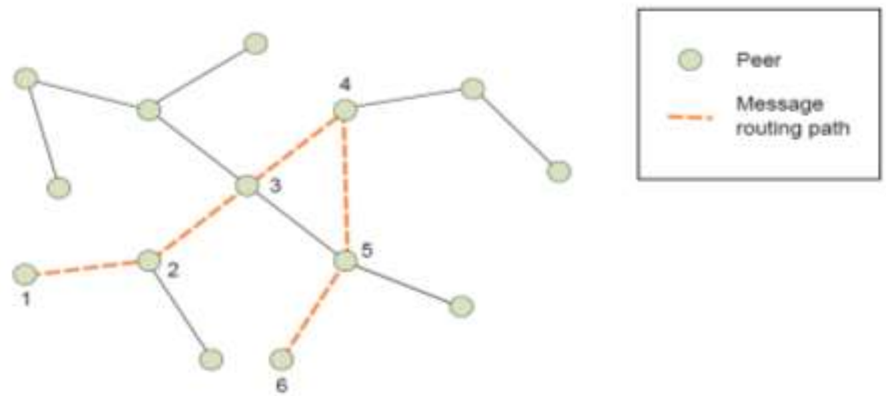
Swarm Intelligence (2)

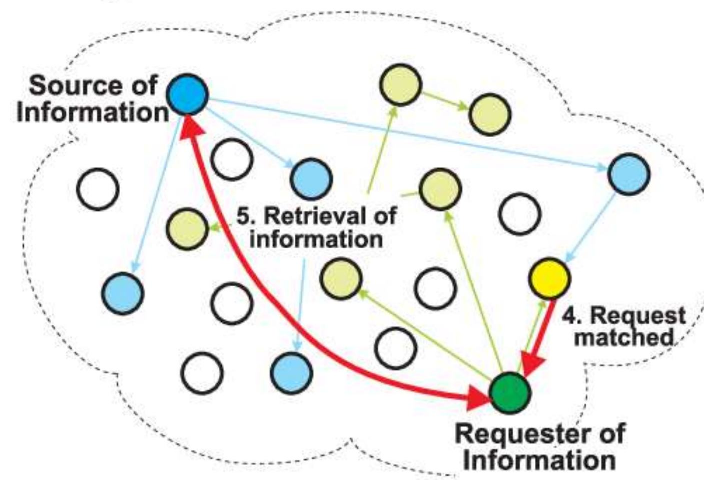
- Individual vs. collective behavior [6] :
  - the individuals within a swarm interact to solve a global objective in a more efficient manner than one single individual could
- Computational study of swarm intelligence, individual organisms studied include *birds* (in flocks), *fishes* (in schools), *ants*, *bees, …*
- Successful applications – e.g., function optimization, finding optimal routes, scheduling, structural optimization, power system controller designs, image and data analysis, etc.

Case Study:
Location and Retrieval of Information in the Internet

Information Placement and Retrieval

- **Information Placement and Retrieval** [6] - data placement and retrieval in fully decentralized P2P networks
  - manipulation of complex, and particularly, incomplete data



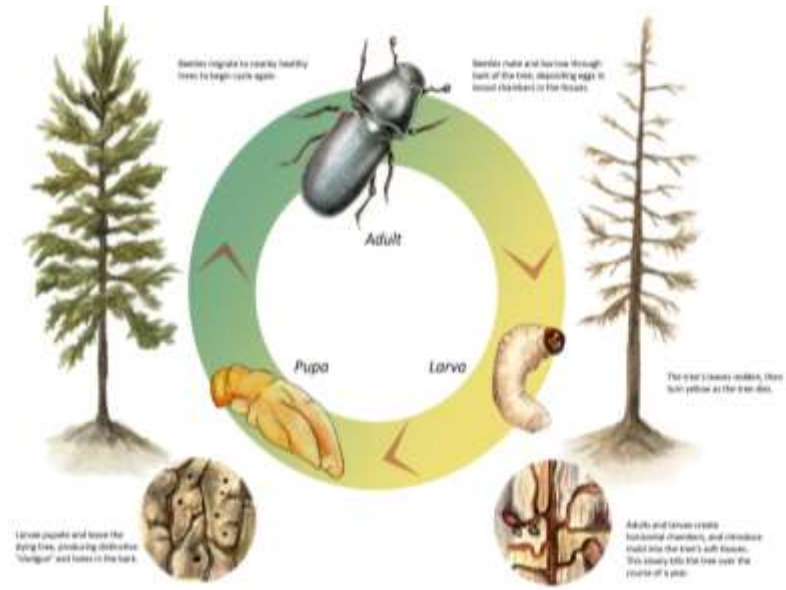


first_name	last_name	birth_date	ID_number	hair_color	height
John				brown	180

- Algorithms applied: Bark Beetles, Slime Mold, Bee Algorithm (BCO), Ant Algorithms (MMAS, AntNet), Gnutella, k-Walker

Bark Beetles

- Bark beetles - bore tunnels called galleries under the bark for feeding and egg laying.
- The life cycle of the bark beetle consists of the following phases [5]:
  - The **reproduction** begins when mature insects arrive on their host tree, where they construct vertical galleries in which they feed and the female beetles deposit eggs.
  - As soon as the larvae have hatched from their eggs, they start feeding and complete their **development** in the galleries.
  - Once **matured**, the beetles **disperse** to search a suitable host for reproduction.
- Bark beetles communicate through two kinds of pheromones:
  - The **attractant pheromone** serves to attract additional beetles to the tree and overwhelm its defenses.
  - The **anti-attractant pheromone** serves to prevent a single tree from being overpopulated.



Bark Beetles life-cycle [8]

P2P Resource Definition (1)

- A resource is defined as a combination of content and its meta-data.
- For simplicity, let us suppose that if two resources have the same content, then their meta-data is equal as well, and vice versa [7].
- Let $S = \{D_1, D_2, ... D_k\}$ be a set of k sets, $k \in N$, which contain values of different data types.
- Each set in S contains a zero element, which is denoted as **nil**.
- A resource is defined as an ordered n-tuple $r = (r_1, r_2, ..., r_n)$, $n \in N$, $k \leq n$, $r \in D_{i_1} \times D_{i_2} \times ... \times D_{i_n}$, where $D_{i_j} \in S$, and $r_i \neq$ **nil**.
- The sets in S define the data type of each element in the resource, and the constraints imposed on it.
- A search query is modeled in exactly the same way, with the only exception that each r_i can be **nil**, but not all of them together.

P2P Resource Definition (2)

- Let $r = (r_1, r_2, \ldots, r_n)$, $n \in N$, $r_i \in D_i$, is a resource.
- The similarity function δ_i is defined as $\delta_i : D_i \times D_i \to R$ and $(\forall r_i)\ \delta_i(\mathbf{nil}, r_i) = 0$. The function itself is normalized.
- The similarity function between queries and resources averages the similarity functions associated with each element in the resource:

$$f(q, r) = \frac{\sum_{i=1}^{n} \delta_i(q_i, r_i)}{n}$$

where $q = (q_1, q_2, \ldots q_n)$ is a query, $r = (r_1, r_2, \ldots r_n)$ is a resource, $n \in N$, and δ_i is a similarity function, defined for position i.

For each query q, an objective function is mapped as $f_q(r) = f(q, r)$ where $r = (r_1, r_2, \ldots r_n)$, $n \in N$ is a resource, and f is the similarity function. The goal of the lookup mechanism is to minimize f_q, which means to find such peer node in the network, having a resource r, that $f_q(r)$ is the minimum.

Bark Beetle Algorithm for Unstructured P2P Search - BB-P2P

CONFIGURABLE PARAMETERS IN THE BEETLE MOVEMENT

Parameter	Description
sufficient_pheromone_concentration	The sufficient pheromone concentration to move along the gradient
too_sufficient_pheromone_concentration	The pheromone concentration when the anti attractant pheromone is emitted
pheromone_radius	TTL of the pheromone messag
attractant_pheromone	The pheromone concentration that is emitted, if a result is found
anti_attractant_pheromone	The pheromone concentration that is emitted, if the detected pheromone concentration is too high

Algorithm Movement of bark beetle

if TTL ≤ 0
then return to source node and terminate search;
else
  if current node has resource matching the search query
  then
    flood neighbors with attractant_pheromone
    return to source node with found resource;
  else
    if the current node has no neighbors
    then return to source node and terminate search;
    else
      if a neighbor has a pheromone concentration $\geq$ too_sufficient_pheromone
      then flood neighborhood with anti_attractant_pheromone ,
      choose the neighbor with highest pheromone concentration;
      if highest pheromone concentration $\geq$ sufficient_pheromone_concentration
      then move to neighbor with highest pheromone concentration;
      else choose a random neighbor and move to it;

Evaluation & Comparison

- Benchmarking of algorithms: Bark Beetle BB-P2P, Physarum Polycephalum PP-P2P, Slime Mold SMP2P, AntNet, Gnutella, and k-Walker [5].
- The analysis encompasses the following stages:
  - First, the parameter sensitivity analysis is done, because of many configurable parameters;
  - Second, the parameters which showed the best results are chosen for the comparative analysis.

Simulation Methodology (1)

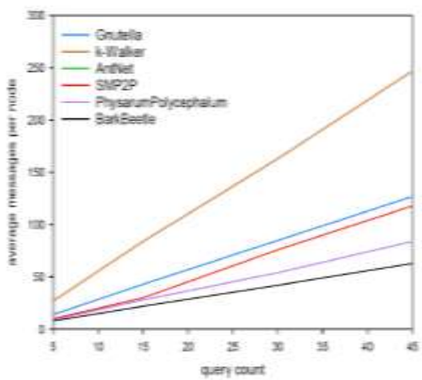
- The P2P graph:
  - Three different network sizes are used for the benchmarking: 50 nodes, 100 nodes and 200 nodes.
- The query distribution is defined as the number of queries sent into the P2P network.
  - To represent different load scenarios, the number of queries are distinguished into four groups relative to the network size, covering a span of low load to very high load: 10%, 30%, 60%, 90%.
- The replication defines the number of nodes in the P2P network, where a specific resource resides.
  - The replication ratio of a specific resource is relative to the network size.
  - Two different replication strategies: *) only one node has the resource and *) the resource is distributed to 16% of the number of nodes of the network (to increase the probability of success).

Simulation Methodology (2)

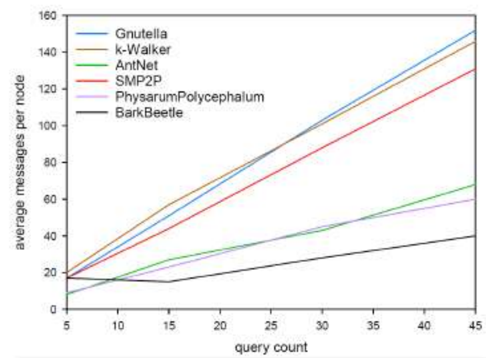
- The following three metrics are used:
  - **Percentage of successful queries:**
  - A query is successful, if it returns exact or acceptable data.
  - **Average messages per node:**
  - It refers to the average message load for each node in the network.
  - **Absolute time:**
  - It represents the elapsed time in milliseconds for a query to be resolved.
- Environment:
  - All simulations are performed in the Google Compute Engine cloud infrastructure 4.
  - For this, a "n1-standard-16" instance is used with Ubuntu 17.10 as the operating system.
  - The instance includes 16 vCPUs and 60GB RAM.

Competitive Analysis

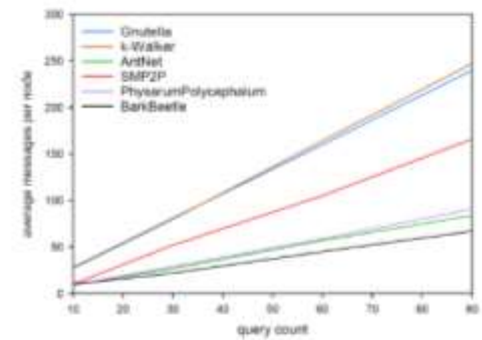
One of metrics implemented to present results:
average message per node



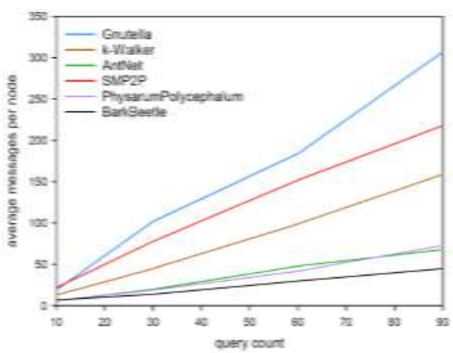
) Peer Model average message per node: 50 nodes, 2% repl



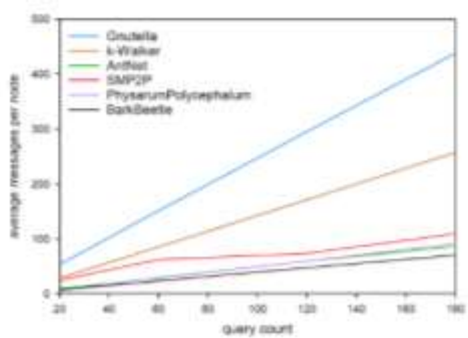
) Peer Model average message per node: 50 nodes, 16% repl.



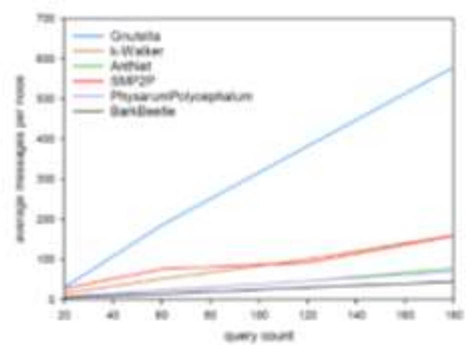
I Peer Model average message per node: 100 nodes, 1% repl.



l) Peer Model average message per node: 100 nodes, 16% repl



) Peer Model average message per node: 200 nodes, 0.5% repl.



/ Peer Model average message per node: 200 nodes, 16% repl.

Conclusion

- The advantages of using bark beetle intelligence in the context of information retrieval in unstructured P2P systems
- The advantages of using intelligent lookup mechanisms in searching, locating and retrieving information
- The advantages of using swarm intelligence in the combination with the other algorithms (both intelligent and unintelligent) in solving information retrieval problem in more complex network structures

Selected References

[1] Sayama H. (2010) Collective Dynamics of Complex Systems, Research Group at Binghamton University, State University of New York

[2] Camazine S., Deneubourg J., Franks N.R., Sneyd J., Theraulaz G. and Bonabeau E. (2003) Self-Organization in Biological Systems, Princeton University Press.

[3] Heylighen F. (2001) The Science of Self-Organization and Adaptivity, The Encyclopedia of Life Support Systems, EOLSS Publishers, Oxford.

[4] https://phys.org/news/2014-05-defense-cyberattacks-power-grids.html

[5] Šešum-Čavić V., Kühn E., Fleischhacker L. Efficient Search and Lookup in Unstructured P2P Overlay Networks inspired by Swarm Intelligence, IEEE Transactions on Emerging Topics in Computational Intelligence, 4(3):351-368, 2020.

[6] Šešum-Čavić V. Swarm Intelligence in Distributed Systems Use-cases, Keynote Lecture, 11th International Joint Conference on Computational Intelligence IJCCI, 2019.

[7] Šešum-Čavić V., Kühn E., Kanev D. (2016) Bio-Inspired Search Algorithms for Unstructured P2P Overlay Networks, Swarm and Evolutionary Computation, 29:73-93, Elsevier.

[8] https://violetsyzhao.weebly.com/beetles-and-hosts.html