Future Generation Computer Systems 112 (2020) 1070–1092

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Optimizing sensors placement in complex networks for localization of hidden signal source: A review

Robert Paluch [a], Łukasz G. Gajewski [a], Janusz A. Hołyst [a,b,*], Boleslaw K. Szymanski [c,d]

[a] *Warsaw University of Technology, Poland*
[b] *ITMO University, Sankt Petersburg, Russia*
[c] *Rensselaer Polytechnic Institute, USA*
[d] *Wroclaw University of Science and Technology, Poland*

ARTICLE INFO

Article history:
Received 19 February 2020
Received in revised form 25 May 2020
Accepted 15 June 2020
Available online 22 June 2020

Keywords:
Information diffusion
Sensors placement
Source localization
Node selection strategies
Network science
Online social networks

ABSTRACT

As the world becomes more and more interconnected, our everyday objects become part of the Internet of Things, and our lives get more and more mirrored in virtual reality, where every piece of information, including misinformation, fake news and malware, can spread very fast practically anonymously. To suppress such uncontrolled spread, efficient computer systems and algorithms capable to track down such malicious information spread have to be developed. Currently, the most effective methods for source localization are based on sensors which provide the times at which they detect the spread. We investigate the problem of the optimal placement of such sensors in complex networks and propose a new graph measure, called Collective Betweenness, which we compare against four other metrics. Extensive numerical tests are performed on different types of complex networks over the wide ranges of densities of sensors and stochasticities of signal. In these tests, we discovered clear difference in comparative performance of the investigated optimal placement methods between real or scale-free synthetic networks versus narrow degree distribution networks. The former have a clear region for any given method's dominance in contrast to the latter where the performance maps are less homogeneous. We find that while choosing the best method is very network and spread dependent, there are two methods that consistently stand out. High Variance Observers seem to do very well for spread with low stochasticity whereas Collective Betweenness, introduced in this paper, thrives when the spread is highly unpredictable.

1. Introduction

The recent development of consumer electronics, social media and online services is changing our society with unprecedented pace. The Internet is used no longer only for communication but also for shopping, banking, learning, entertainment, and most of all, sharing and searching for information. The number of smart devices used in everyday life is growing, most of them connected to the Internet to access Online Social Networks (OSN), hosted by such platforms as Twitter, Facebook, or Instagram. The Internet became the primary source of news, opinions and comments for many people [1]. As always, new technologies bring not only new opportunities but also new threats. The Internet has become a target for wide range of attacks, including propagation of misinformation, and malware, as well as medium for fraudulent activities [2,3]. Naturally, such nefarious spread is on fundamental level similar to traditional epidemiology concerned with biological pathogen spread and therefore those two are often compared [4–12]. While the process of spreading of biological or digital viruses has been well studied for many years [13–16], new methods for detecting and preventing malicious content in OSN are still being developed [17,18]. The pace of this research has been accelerating recently, particularly in regard of an important problem of locating spread source of malicious or harmful information [19–29].

Despite all these efforts, many challenges still remain. One such challenge is an optimal sensor placement to monitor the network of interest in the least expensive and unobtrusive way. There have been plenty of recent works tackling this issue (see Section 2 below), however, there is a lack of comprehensive, state of the art, comparative studies of this research. This motivated us to conduct such a review ourselves. In this paper, we compare performance of six algorithms (five representing the best known solutions and the sixth used as a null model, see Section 4 below) used in this field in various scenarios with varying spread stochasticity, network topology and the amount of information accessible by the spread monitoring sensors.

[*] Corresponding author at: Warsaw University of Technology, Poland.
E-mail address: janusz.holyst@pw.edu.pl (J.A. Hołyst).

https://doi.org/10.1016/j.future.2020.06.023
0167-739X/© 2020 Elsevier B.V. All rights reserved.

Table 1
Computational complexity of studied algorithms. Here $n = |V|$ is the number of nodes, $m = |E|$ is the number of links, $b = |S|$ is the number of sensors (budget), and γ is an experimentally obtained scaling exponent from the linear model best fitting to results of numerical simulations (see Fig. 1).

Algorithm	Complexity	γ	Parallelizable?
High Coverage Rate [30]	$O(bn)$	1.80(2)	No
Betweenness Centrality [31]	$O(nm)$	2.17(1)	Yes
Collective Betweenness	$O(nm + bn\log^2 n)$	2.77(3)	Partially
High Variance Observers [32]	$O(b^2n^2)$	3.91(1)	Yes
K-Median [33]	$O(b^2n^2)$	4.08(3)	Yes

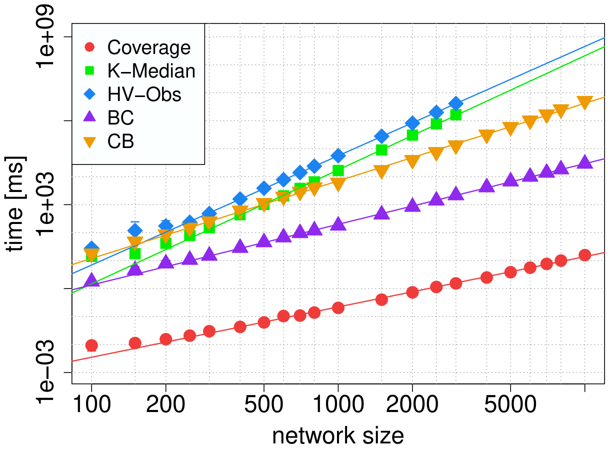

Fig. 1. Average time needed for determining the sensors set (density of sensors $\rho = 0.05$) versus the size of Erdős–Rényi network with average degree $\langle k\rangle = 8$. Error bars representing confidence interval of level 0.95 are smaller than symbols in most cases. Solid lines are linear models $\ln(\text{time}) = \gamma \ln(\text{size}) + \text{const}$. The values of γ are presented in Table 1. HV-Obs and K-Median were parallelized using 16 threads.

2. Related works

Spinelli & Celis et al. have introduced a method called High Variance Observers (Section 4.4) and tested it in the context of what they call a *budget* and a transmission variance [32,34,35]. The budget is simply the number of sensors we are allowed to use while the transmission variance is, of course, a measure of how non-deterministic the monitored spread is. While their results look promising, we find both the range of transmission variance and the number of networks considered not comprehensive enough.

Zhang et al. have analyzed centrality based methods and showed that none is be a clear-cut winner in terms of performance [30]. While in their work the range of sensor densities is respectable, they do not consider effects of stochasticity at all. Based on their results, for our comparisons, we chose one of the centrality measures (Section 4.1) that seems slightly better than the rest of them.

In contrast to above work, we consider both – sensor density and transmission variance – in a wide range of values and on top of that we use eight different networks as our testing environments.

There have also been some works on *online* sensors selection to be able to take into account the spread evolving dynamics [35–37]. However, we consider such approach to be beyond the scope of our study because it introduces new issues and challenges.

On similar note there has been some impressive work by Zejnilovic et al. [38–40], Wang [23], Fang et al. [41], Li et al. [29] and Shi et al. [42] where new localization methods are introduced and several methods for sensor placement are studied. Nevertheless, the localization scheme and, most importantly, certain assumptions are different than the ones we are using in this paper and therefore we will omit those as well.

3. Basics

3.1. Spreading model

We use Susceptible–Infected model [43] to simulate propagation over the complex network. SI model has only one parameter, *infection rate* β, which is a probability per time step that infected node will transmit the infection to the uninfected neighbor. The infected nodes try to infect their neighbors in every time step. The distribution of the number of time steps needed to transmit the infection is geometric, with $\mu = 1/\beta$ and $\sigma = \sqrt{1-\beta}/\beta$. We define *transmission variance* ξ according to Spinelli et al. [32], as the ratio between the standard deviation and the mean of the number of times steps needed to transmit the infection $\xi = \sigma/\mu = \sqrt{1-\beta}$. Please note that ξ is nothing else than inverse of propagation ratio λ [25,28].

3.2. Source localization algorithm

For the localization of spread source, Pinto-Thiran-Vetterli [25] algorithm is used in restricted form (PTVA-LI [28]).

Let us have a network $G = (V, E)$, with V, E being its known sets of vertices (nodes) and edges (links) respectively, defining our system. In this system, we place our set of sensors $S \subset V$, that is the nodes that report at what time they got infected, and an unknown origin of infection o^*. Normally (i.e., in PTVA), we would also register from whom given sensor received the virus, however, this is where the *restricted form* part comes in — we use *limited information* version, namely PTVA-LI and only the times of infection are given by the sensors.

We assume that the inception time of the virus t_0 is unknown, and only the mean and variance of the transmission time μ, σ per link are known (but not exact propagation times).

The goal, of course, is to locate o^*

From infection times reported by sensors we construct an observed delay vector $\mathbf{d}$:

$$\mathbf{d} = (t_2 - t_1, t_3 - t_1, \ldots, t_b - t_1)^T \qquad (1)$$

where b is the number of sensors (budget), t_i is an infection time of sensor $s_i \in S$, and t_1 is the infection time of a *reference sensor* that is needed here since the t_0 is unknown.

For each node in the system $v \in V$ we compute a deterministic delay vector $\boldsymbol{\mu}$:

$$\boldsymbol{\mu}_v = \mu\big(|P(v, s_2)| - |P(v, s_1)|, \ldots, |P(v, s_b)| - |P(v, o_1)|\big)^T \qquad (2)$$

where $|P(v, s_i)|$ is number of edges on a shortest path connecting nodes v, s_i. We also compute the covariance matrix $\boldsymbol{\Lambda}_v$, each element i, j of which is given by:

$$\boldsymbol{\Lambda}_{i,j} = \sigma^2 \times \begin{cases} |P(s_i, s_1)| & i = j, \\ |P(s_i, s_1) \cap P(s_j, s_1)| & i \neq j \end{cases} \qquad (3)$$

Finally we compute a *score* for each node v and use maximum likelihood rule to determine the most probable origin of the epidemic $\hat{o}$:

$$\hat{o} = \underset{v\in V}{\arg\max}\, \boldsymbol{\mu}_v^T \boldsymbol{\Lambda}_v^{-1}(\mathbf{d} - 0.5\boldsymbol{\mu}_v) \qquad (4)$$

When $\hat{o} = o^*$ we count it is as a success. See evaluation metrics below for details.

Since in general G can be any graph and PTVA is optimal on trees, one must construct a BFS (breadth first search) tree on each node $v \in V$ and apply the above described procedure (eq. (2)–(4)) on each tree respectively.

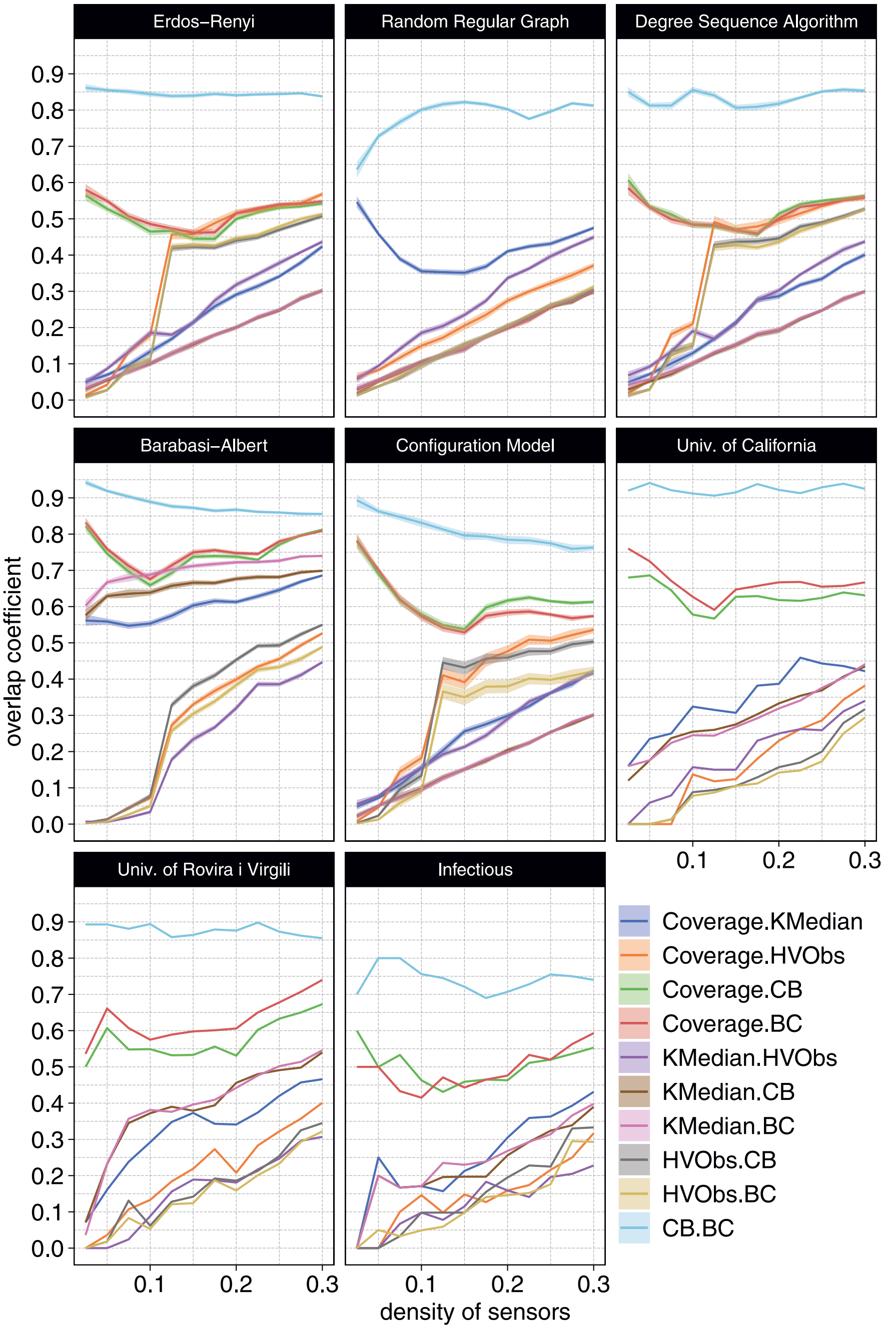

Fig. 2. The overlap coefficient as a function of density of sensors for all networks used in this study. All panels have the same horizontal axis. The results for synthetic networks are averaged over 50–150 realizations. Error bands represent the confidence intervals at the level 0.95.

4. Algorithms

We compare the following six methods of efficient sensors' selection.

4.1. Betweenness centrality (BC)

This popular and simple heuristic takes the nodes with largest betweenness centrality, which is computed independently for every node $v \in V$ as

$$S_{\mathrm{BC}} = \arg\max_{S} \sum_{v\in S} \sum_{\substack{i,j\in V\\ i\neq j\neq v}} \sigma_{ij}^{(v)}/\sigma_{ij}, \quad (5)$$

where $\sigma_{ij}^{(v)}$ is the number of shortest paths between i and j which contain v and σ_{ij} is the total number of shortest paths between them.

4.2. High coverage rate (coverage)

The algorithm, proposed by Zhang et al. selects a set of the sensors S_{Coverage} which has the maximum number of unique neighbors [30]. The nodes which have an sensor as a neighbor are *covered*, and the fraction of covered nodes in the network is called *coverage rate*. This method saturates, since usually the density of sensors which gives coverage rate equal one is significantly smaller than one.

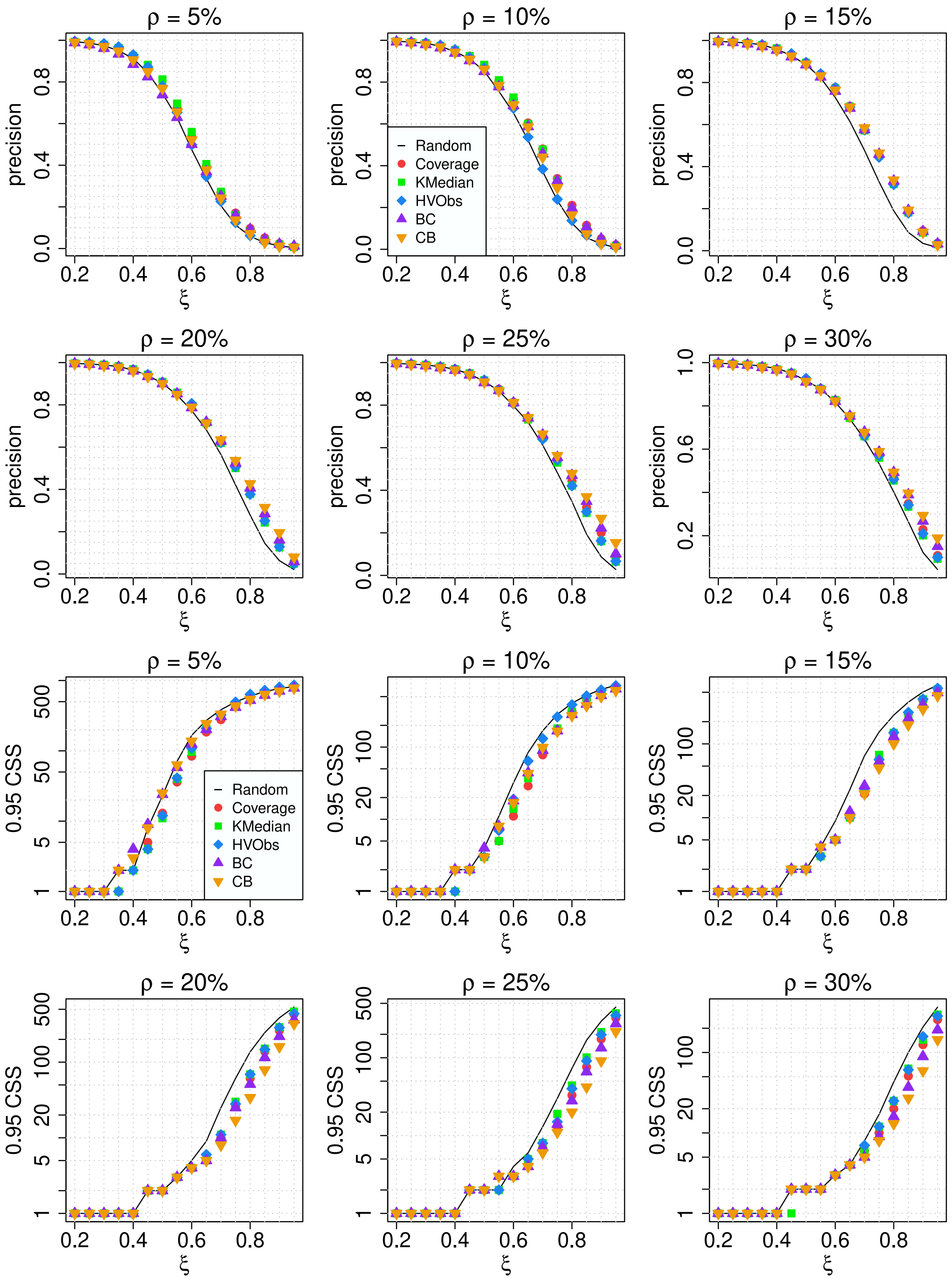

Fig. 3. Comparison of source localization quality for five sensor placement strategies in case of **Erdős–Rényi graph** ($n = 1000$, $\langle k \rangle = 8$). A black solid line denotes the results for randomly placed sensors. The confidence intervals at the level 0.95 are smaller than symbols.

To avoid the problem of saturation, additional stages are added to the algorithm in the following manner. The first stage is identical as originally proposed by Zhang et al. [30]. The sensors are chosen greedily, one by one, and each new sensor increases the coverage rate until it reaches unity. When the coverage rate became one, the next stage starts. In the second stage, the algorithm selects sensors which maximize the number of nodes which have two sensors as the neighbors (*double-covered* nodes). In the third stage algorithm maximizes the number of *triple-covered* nodes and so on until the desirable density of sensors is reached.

4.3. K-median (K-Median)

K-Median placement was proposed by Berry et al. for efficient detectability of a flow in municipal water networks [44]. This method minimizes the sum of distances between the nodes and

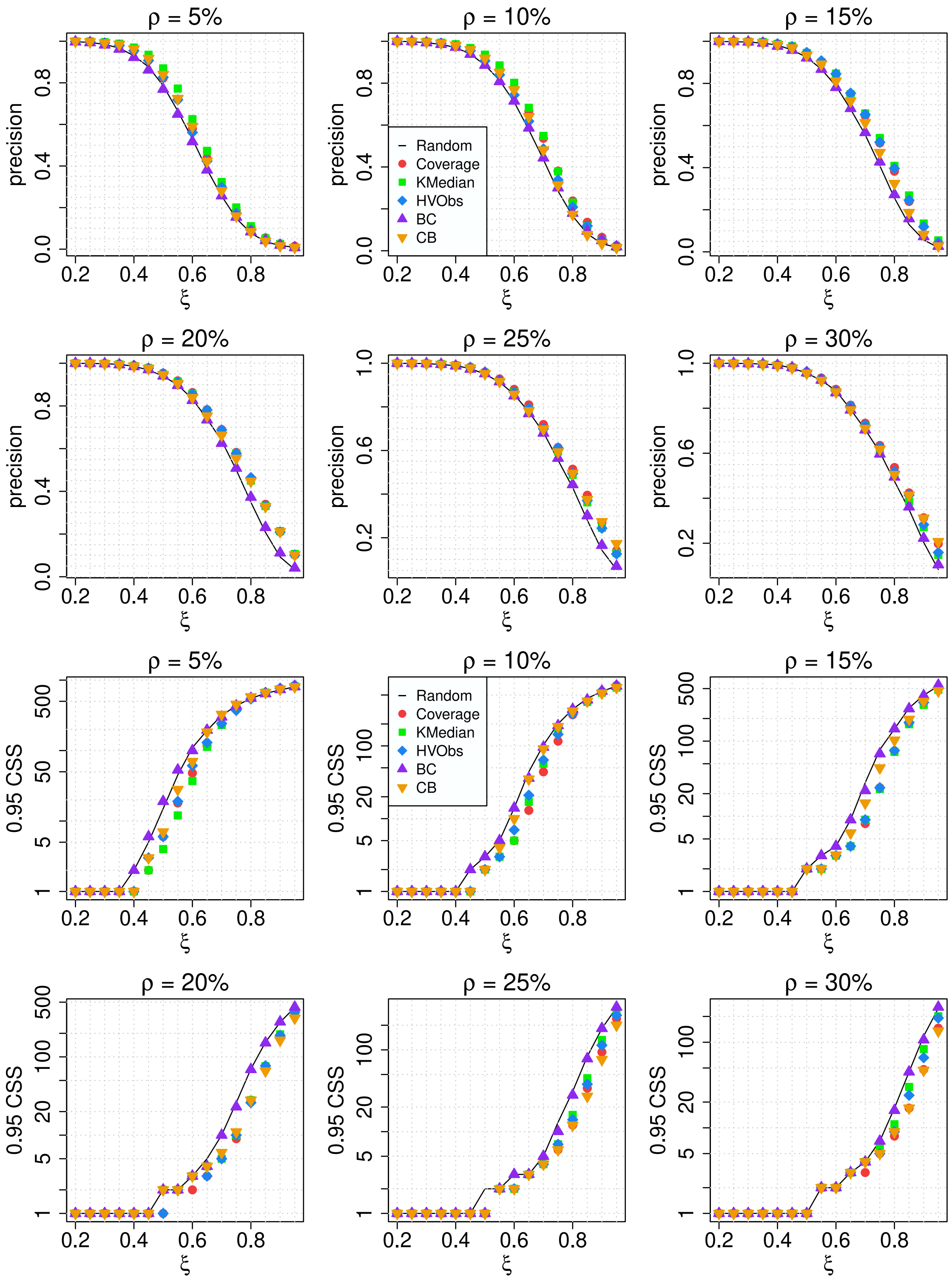

Fig. 4. Comparison of source localization quality for five sensor placement strategies in case of **Random Regular Graph** ($n = 1000$, $\langle k \rangle = 8$). A black solid line denotes the results for randomly placed sensors. The confidence intervals at the level 0.95 are smaller than symbols.

the closest sensors. If V is the set of all nodes, $d(i,j)$ is the length of the shortest path between nodes i and j, $S_{\text{K-Median}}$ set of sensors is

$$S_{\text{K-Median}} = \arg\min_{S} \sum_{i \in V} \min_{o \in S} d(i,o) \tag{6}$$

4.4. High variance observers (HV-Obs)

The algorithm introduced by Spinelli et al. is based on path covering strategy [32]. This method looks for a set of sensors $S_{\text{HV-Obs}(L)}$ that maximizes the cardinality of $P_L(S)$, which is the set of nodes that lie on a shortest paths of length at most L between any two sensors in the set S.

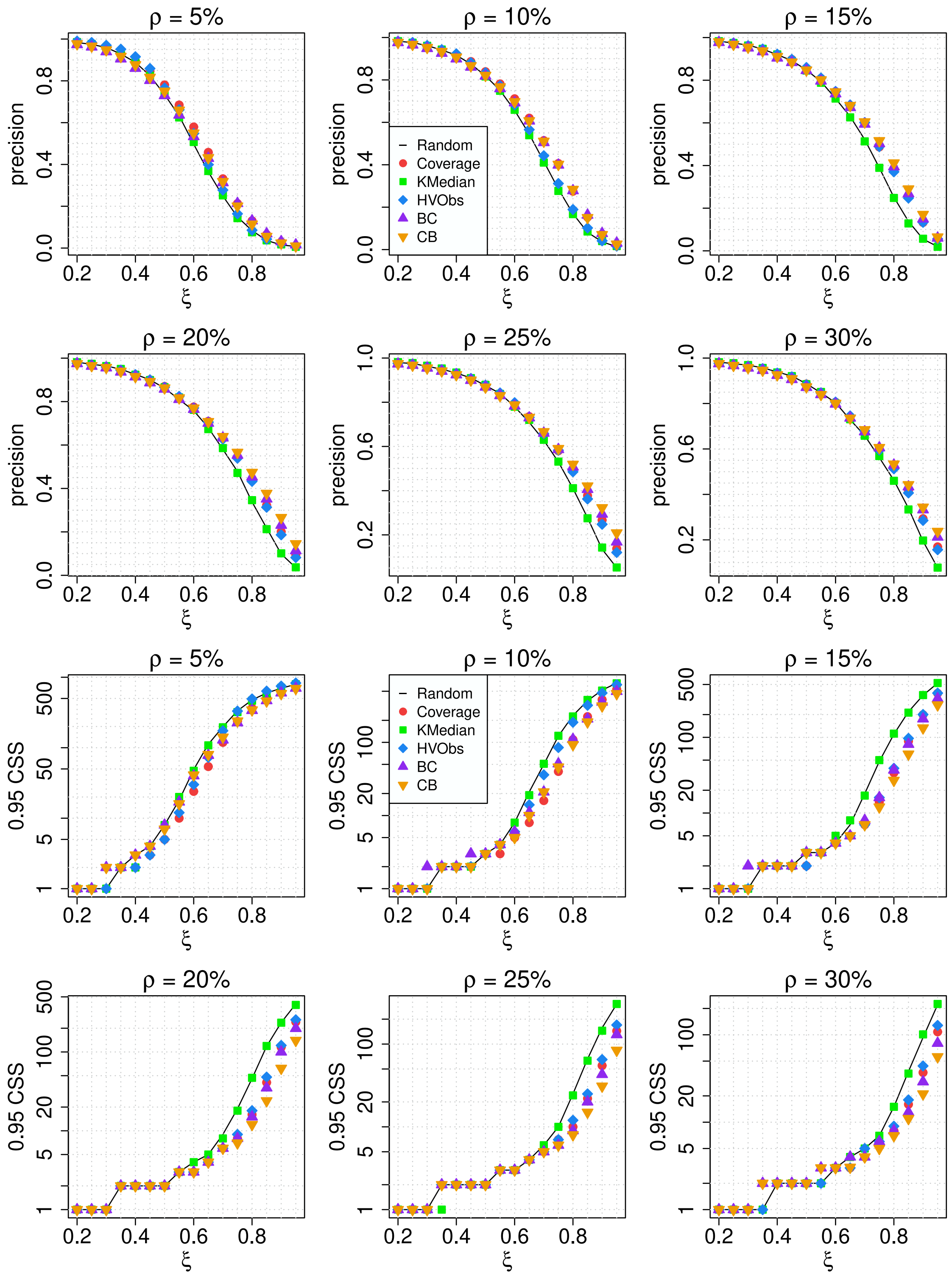

Fig. 5. Comparison of source localization quality for five sensor placement strategies in case of **Degree Sequence Algorithm** ($n = 1000$, $\langle k \rangle = 7.67$). A black solid line denotes the results for randomly placed sensors. The confidence intervals at the level 0.95 are smaller than symbols.

High Variance Observers is designed for small density of sensors, since $|P_L(S)|$ increases quickly with the number of sensors and can easily reach the number of all nodes in the network. Here, to solve this problem, we extend the algorithm in the similar way as in the case of High Coverage Rate method. A node which lies on an exactly one shortest path of length at most L between any two sensors in the set S is called *single-path-covered*. A *double-path-covered* node lies on two shortest paths, *triple-path-covered* lies on three shortest paths and so on. In the first stage, the algorithm selects greedily the sensors which maximize the number of *single-path-covered* nodes until all nodes are *single-path-covered*. Then, the second stage starts, in which the number of *double-path-covered* is maximized and so on until the desirable density of sensors is reached.

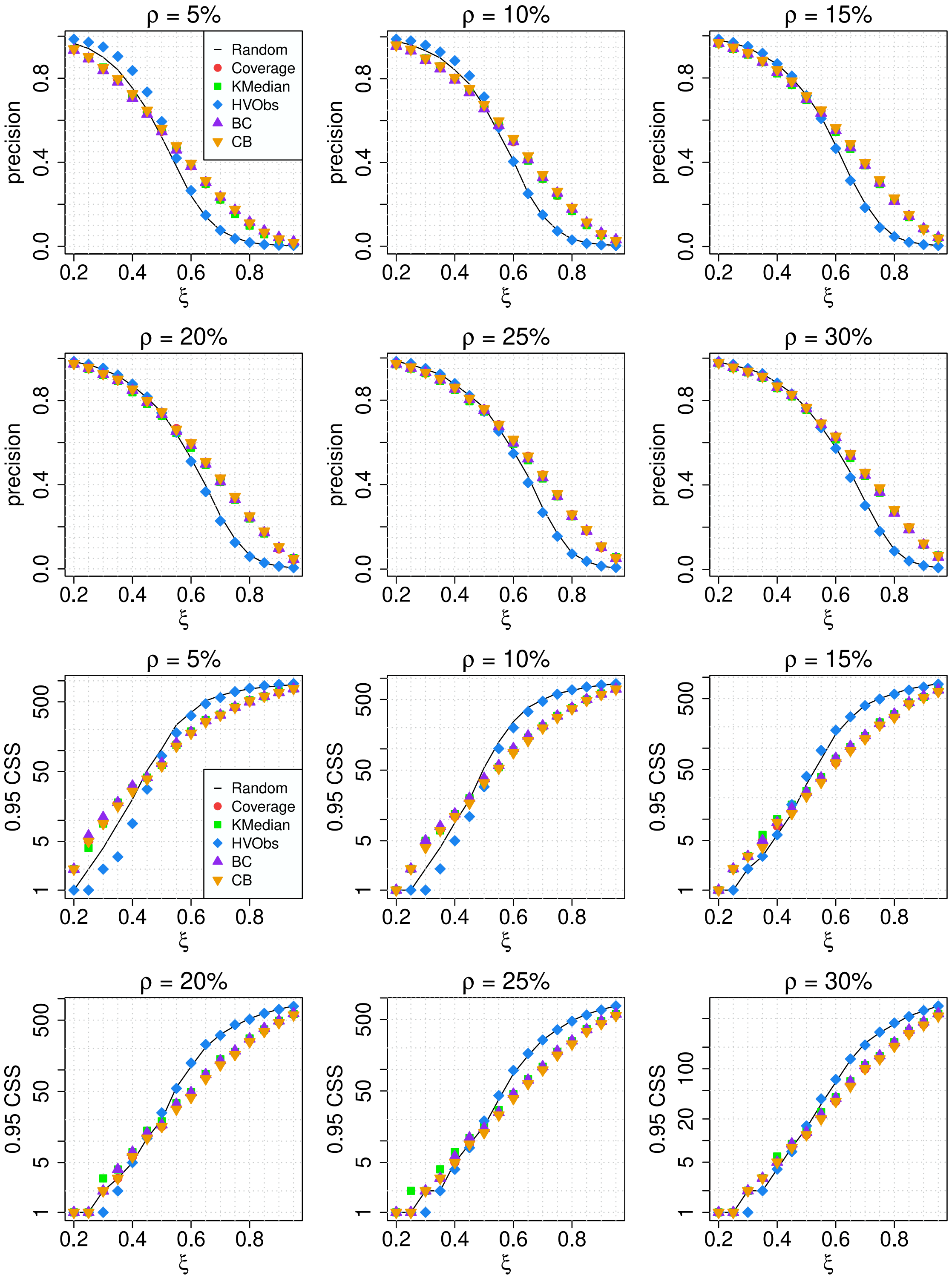

Fig. 6. Comparison of source localization quality for five sensor placement strategies in case of **Barabási–Albert model** ($n = 1000$, $\langle k \rangle = 7.98$). A black solid line denotes the results for randomly placed sensors. The confidence intervals at the level 0.95 are smaller than symbols.

4.5. Collective betweenness (CB)

We propose a novel method which maximizes a new measure called Collective Betweenness:

$$S_{\mathrm{CB}} = \arg\max_{S} \sum_{\substack{i,j \in V \\ i,j \notin S}} \sigma_{ij}^{(S)}/\sigma_{ij}, \qquad (7)$$

where S denotes a set of sensors, $\sigma_{ij}^{(S)}$ is the number of shortest paths between i and j which pass through any node belonging to S and σ_{ij} is the total number of shortest paths between them. In this approach each shortest path is counted only once (even if it passes through many nodes belonging to S), therefore value of CB for the set S is different than simple sum of Betweenness Centralities of all nodes in S.

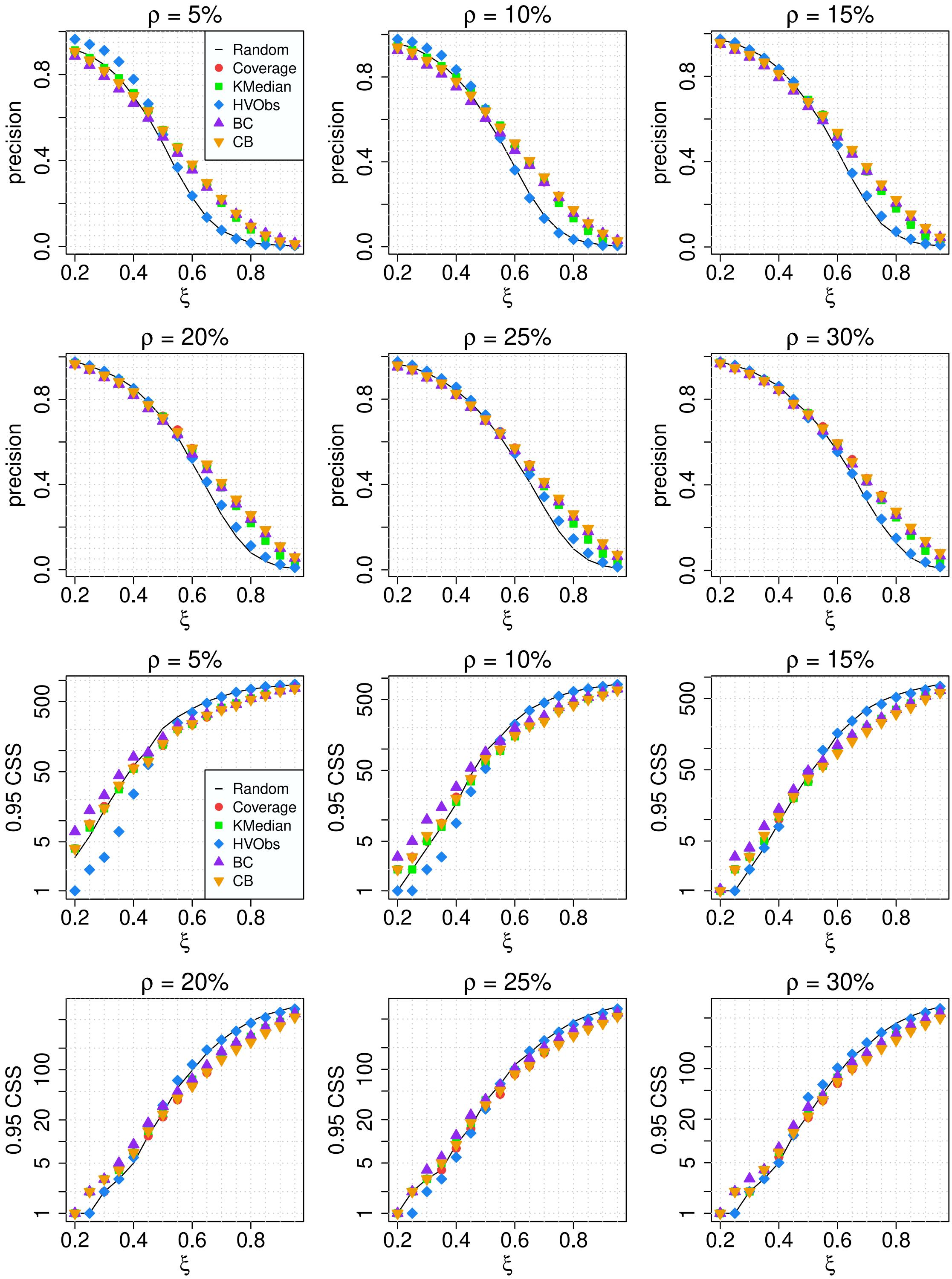

Fig. 7. Comparison of source localization quality for five sensor placement strategies in case of **Configuration Model** ($n = 1000$, $\langle k \rangle = 7.02$). A black solid line denotes the results for randomly placed sensors. The confidence intervals at the level 0.95 are smaller than symbols.

4.6. Random

This is a benchmark for the rest of methods (a baseline). This method selects sensors randomly.

4.7. Complexity

Computational complexity is an important criterion for the usability of the algorithms. As seen in Table 1, time complexity of methods studied in this article varies from $O(nm)$ to $O(b^2n^2)$, where $n = |V|$ is the number of nodes, $m = |E|$ is the number of links and $b = |S|$ is the number of sensors (budget). The third column contains the results of numerical experiment in which the average execution time for each method was measured as a function of network size (see with Fig. 1). The experiment was conducted for Erdős–Rényi network with constant density of sensors $\rho = 0.05$ and average degree $\langle k \rangle = 8$. High Coverage Rate has the lowest computational complexity among the tested

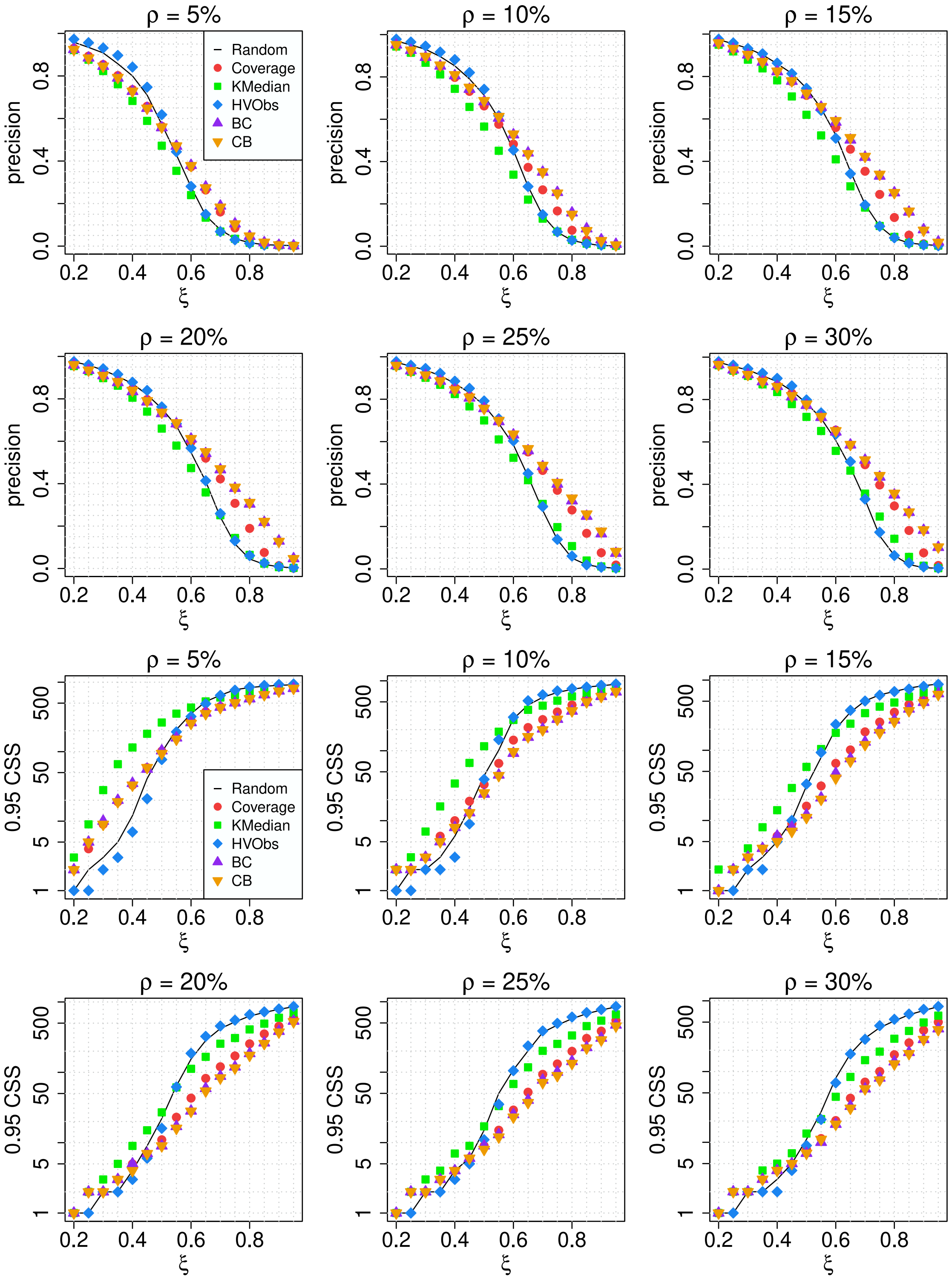

Fig. 8. Comparison of source localization quality for five sensor placement strategies in case of **University of California network** ($n = 1020$, $\langle k \rangle = 12$). A black solid line denotes the results for randomly placed sensors. The confidence intervals at the level 0.95 are smaller than symbols.

methods. The authors state that the complexity of this algorithm can be reduced to $O(n + m)$ [30], but our version, which is resistant to the saturation effect, has complexity $O(bn)$. On the opposite extreme, there are High Variance Observers and K-Median. In case of these algorithms, the maximum network size during the experiment shown in Fig. 1 is 3000 due to very long time of computations. Although both methods are well parallelizable, the maximum possible speed-up is equal to the number of nodes in the network (and requires this number of threads), which can be still insufficient for applying these algorithms to large networks.

4.8. Similarity between sets of sensors

The algorithms for sensor placement are very different (except for BC and CB), yet, the sets of nodes which they generate often are very similar to each other. We use an overlap coefficient [45]

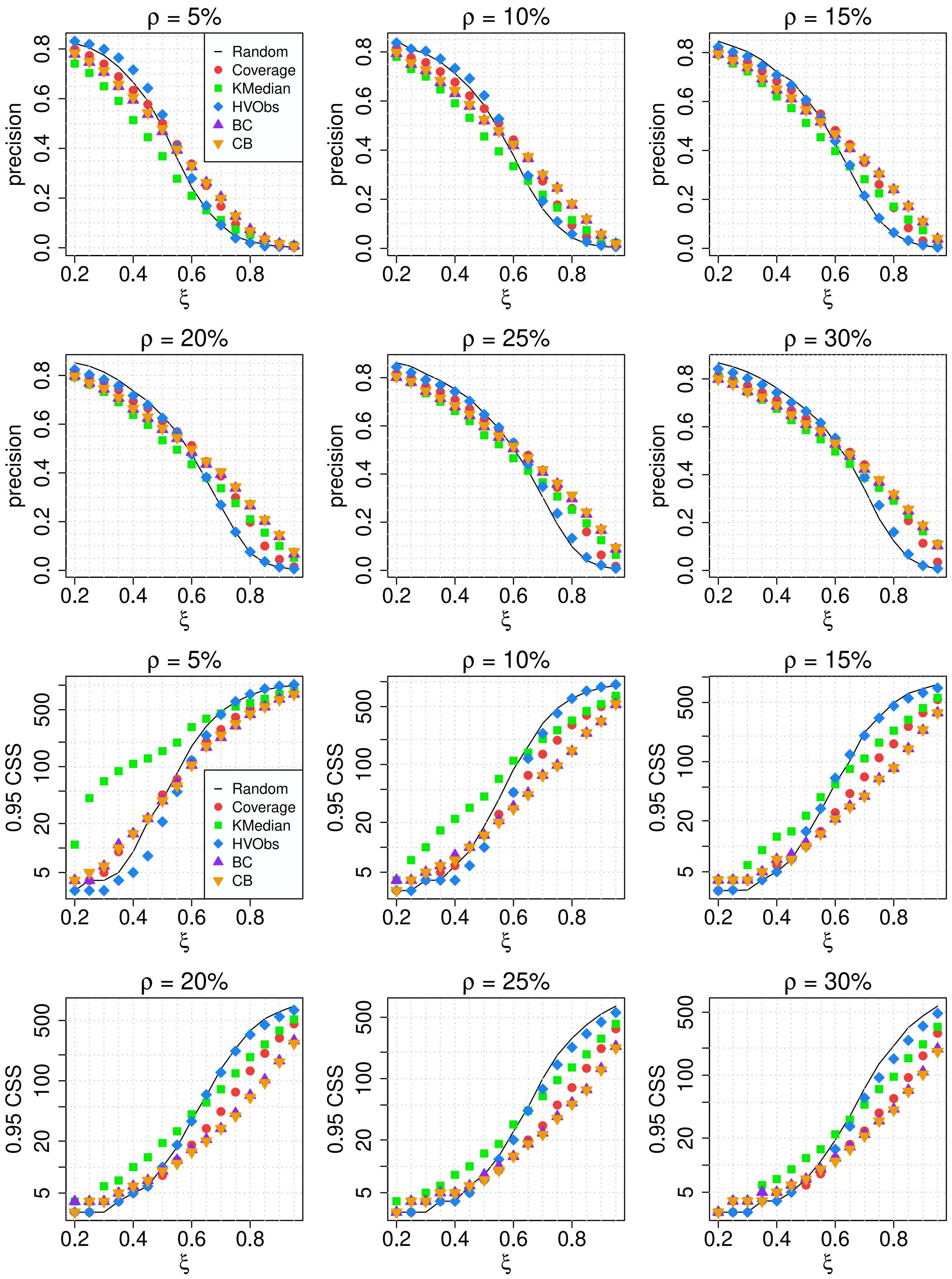

Fig. 9. Comparison of source localization quality for five sensor placement strategies in case of **University of Rovira i Virgili network** ($n = 1133$, $\langle k \rangle = 9.6$). A black solid line denotes the results for randomly placed sensors. The confidence intervals at the level 0.95 are smaller than symbols.

as similarity measure:

$$\text{overlap}(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)} = \frac{|X \cap Y|}{s}, \qquad (8)$$

where $s = |X| = |Y|$ is the number of sensors. Fig. 2 presents the overlap coefficient as a function of the density of sensors for all networks used in this study. The overlap coefficients between random sensors and other sets of sensors are not shown because their values are known and equal to the density of sensors ρ.

As expected, the highest overlap occurs for Betweenness Centrality and Collective Betweenness. It oscillates between 0.7 for the Infectious network and 0.95 for the University of California network. The most unique set of sensors for every network is the set given by High Variance Observers method. Any overlap with this method is always below 0.55.

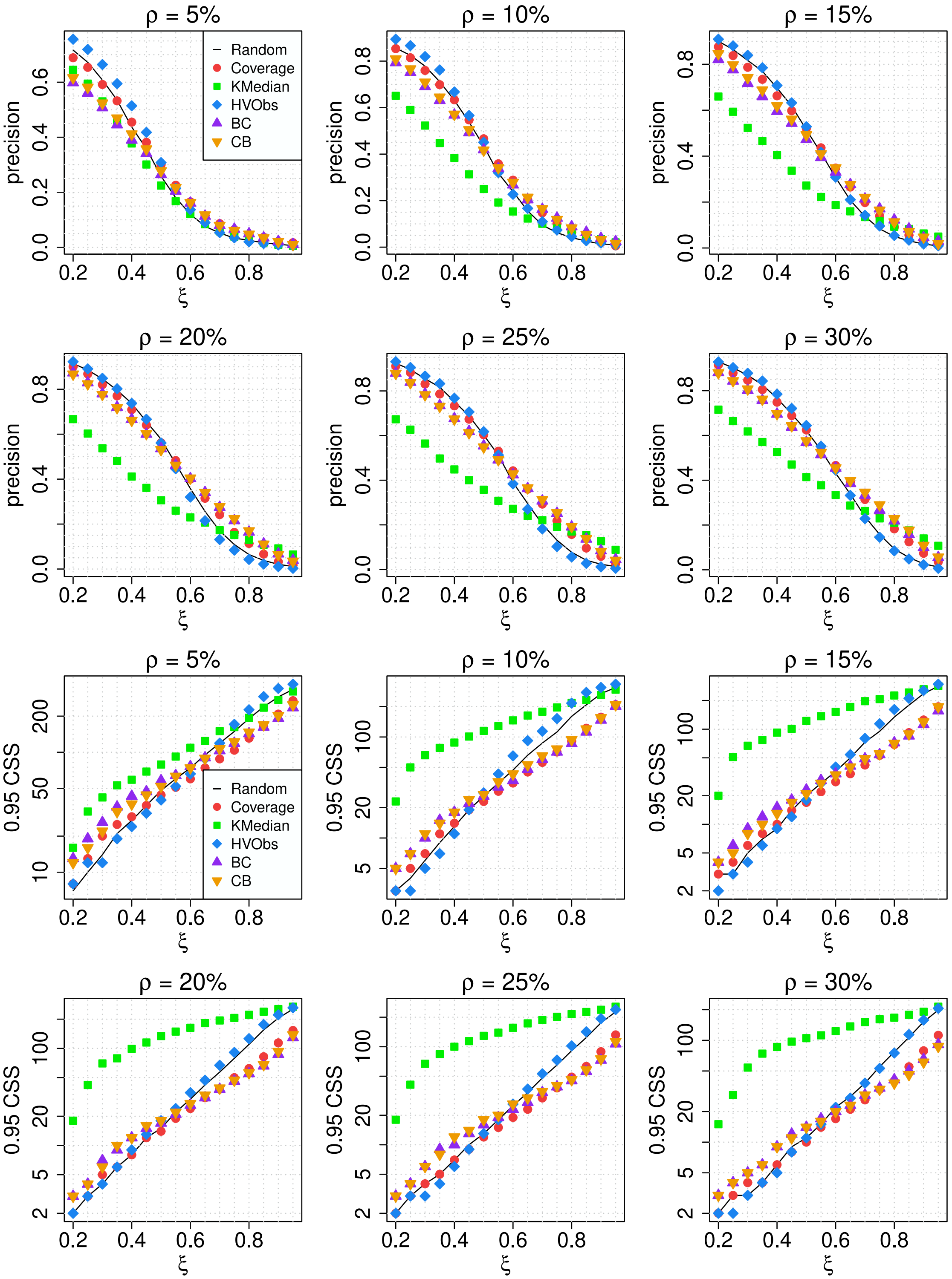

Fig. 10. Comparison of source localization quality for five sensor placement strategies in case of **Infectious network** ($n = 410$, $\langle k \rangle = 13.5$). A black solid line denotes the results for randomly placed sensors. The confidence intervals at the level 0.95 are smaller than symbols.

5. Evaluation metrics

Two efficiency measures are used for evaluating the quality of origin detection: the average precision and the Credible Set Size at 0.95 confidence level. The precision for a single test is defined as the ratio between the number of correctly located sources (i.e., true positives, which here equals either zero or one) and the number of sources found by the method (i.e., true positives plus false positives, which here is at least one). The tests are repeated multiple times for different origins and many graph realizations (for synthetic networks) and then the obtained values of precision are averaged. The Credible Set Size at the confidence level of α (α-CSS) is a novel metric introduced by us here. It is the size of the smallest set of nodes containing the true source with probability

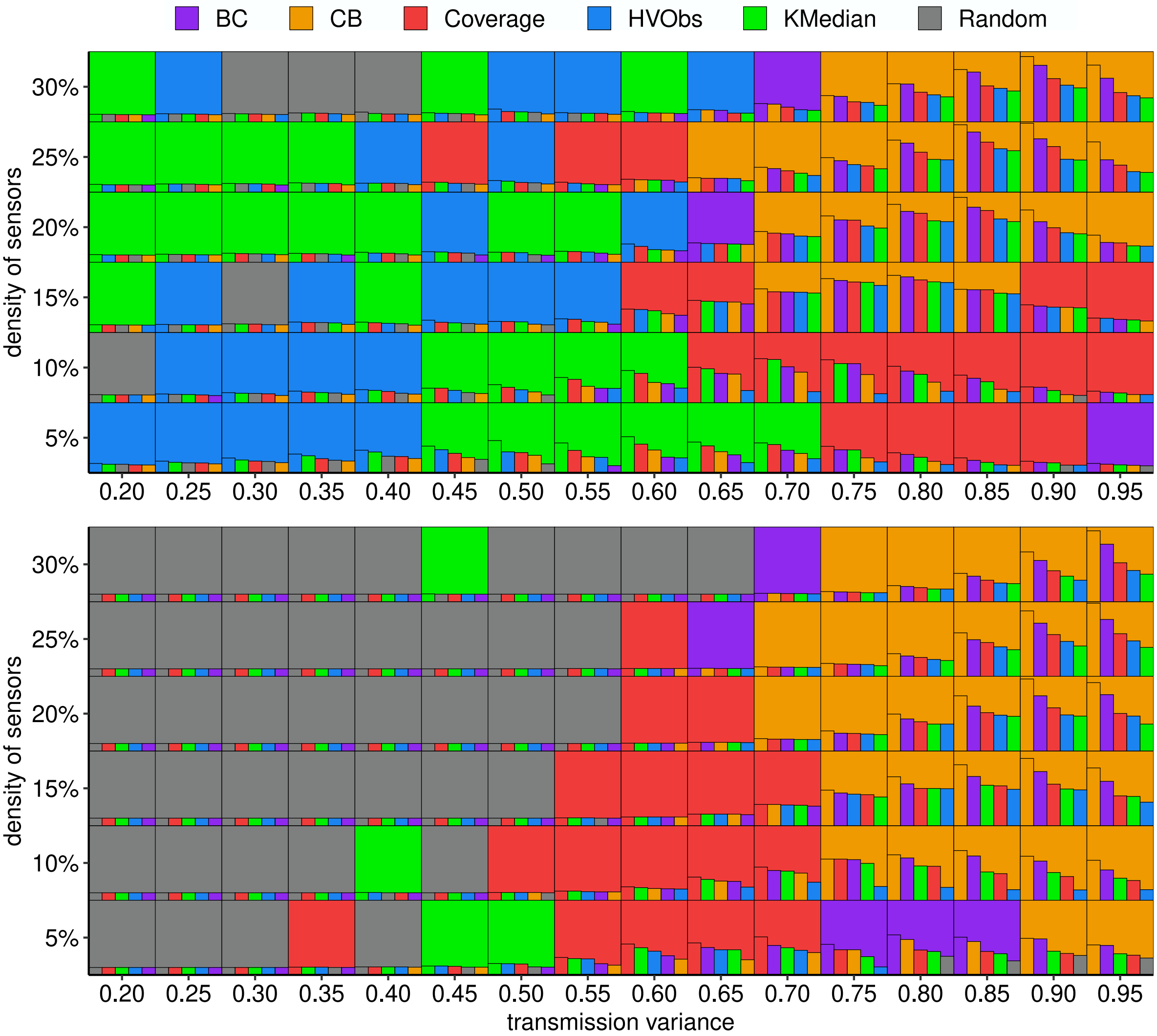

Fig. 11. Summary Diagrams for **Erdős–Rényi** graph ($n = 1000$, $\langle k \rangle = 8$). The color of the background in each tile indicates which algorithm provides the highest average precision (top) or the smallest 0.95-CSS (bottom) for a given pair (ξ, ρ). The five colored bars inside each tile, ordered from highest to lowest, illustrate the ranking of the methods. The height of each bar shows the difference in average precision (or 0.95-CSS) between a given method and the last method. The last (sixth) method, which is the least effective for a given pair (ξ, ρ) is not shown in a given tile. The heights of bars from all tiles are scaled relative to the height of the highest bar among them, with a minimum height to recognize the color. The bars on the left sides of Summary Diagrams (for low transmission variance ξ) are much lower than bars on the right sides, which means that the differences in average precision and 0.95-CSS between methods are much smaller for low values of ξ than for large ξ. In case of ER graph, the highest bar in Summary Diagram for average precision (top, $\xi = 0.9$, $\rho = 25\%$) represents difference of 18(1) percent points, while the highest bar for 0.95-CSS (bottom, $\xi = 0.95$, $\rho = 25\%$) corresponds to 230 nodes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
The number of nodes in each network $n = 1000$. The values of the average degree $\langle k \rangle$, the average path length (APL) and the global clustering coefficient are averages of 100 realizations of graphs.

Graph type	Degree distribution	$\langle k \rangle$	APL	Global clustering
Erdős–Rényi	Binomial	8.00	3.55	0.008
Random Regular Graph	$k = \text{const}$	8.00	3.60	0.006
Degree Sequence Algorithm	Poisson	7.67	3.76	0.132
Barabási–Albert	$P(k) \sim k^{-3}$	7.98	3.17	0.026
Configuration Model	$P(k) \sim k^{-3}$	7.02	3.36	0.021

α. In other words this metric describes how many nodes with the highest *score* should be labeled as origin to have probability α that the true origin is among these nodes. Probability here is understood as frequency and computed as a hit rate (recall) from many realizations of signal propagation and source location.

Table 3
Basic properties of the real networks used in tests.

Network	|V|	$\langle k \rangle$	k_{max}	APL	Diameter	Global clustering
Univ. of California	1020	12.2	110	3.0	5	0.046
Univ. of Rovira i Virgili	1133	9.6	71	3.6	8	0.166
Infectious	410	13.5	50	3.6	9	0.436

6. Results

We evaluate the algorithms for sensor placement on several synthetic networks, such as Erdős–Rényi model [46], Random Regular Graph, Degree Sequence Algorithm [47] with Poisson distribution, Barabási–Albert model [48], Configuration Model with power-law degree distribution and three real networks (one network of human face-to-face contacts and two networks of Internet communications between academics). We study how the

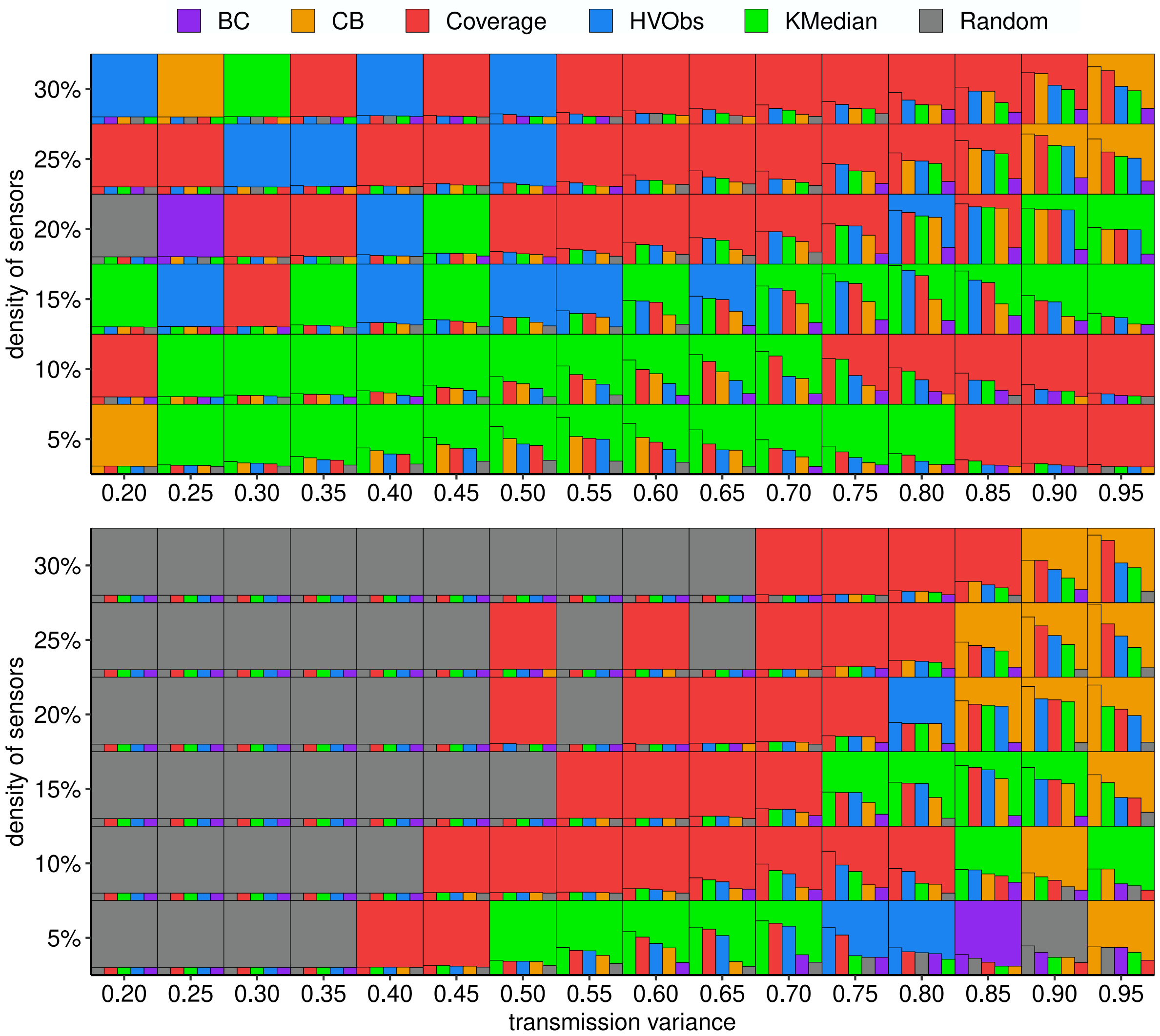

Fig. 12. Summary Diagrams for **Random Regular Graph** ($n = 1000$, $\langle k \rangle = 8$). The highest bar for average precision (top diagram, $\xi = 0.8, \rho = 15\%$) represents difference of 15(1) percent points, while for 0.95-CSS the highest bar (bottom diagram, $\xi = 0.95, \rho = 25\%$) corresponds to 133 nodes. See Fig. 11 for detailed instruction how to read Summary Diagrams. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Summary Table for **Erdős–Rényi** graph ($n = 1000$, $\langle k \rangle = 8$) presents the average values of precision (top numbers in cells, in percentages) and Credible Set Size (bottom numbers in cells) in nine regions of parameter space (ξ, ρ). The first three numerical columns from the left refer to low, three in the middle to medium, and the last three to high transmission variance ξ. Similarly, columns $\{1, 4, 7\}$ correspond to high, $\{2, 5, 8\}$ to middle, and $\{3, 6, 9\}$ to low density of sensors ρ. Due to arrangement of the columns, the average precision (0.95-CSS) always decreases (increases) from left to right side of the table. The best results in each region are printed in bold. The uncertainty of average precision is given by the confidence interval at the level 0.95. In case of ER graph, K-Median, HV-Obs and Coverage are the best methods for low transmission variance ξ, while Collective Betweenness is the leading method for high values of ξ. However, Coverage is also doing well for high ξ if the density of sensors ρ is low.

ξ →	⟨0.2; 0.5⟩			⟨0.5; 0.8⟩			⟨0.8; 0.95⟩		
ρ [%] →	20–30	10–20	5–10	20–30	10–20	5–10	20–30	10–20	5–10
Random	96.9(1) 1	95.9(1) 1	93.3(1) 2	67.3(2) 17	57.3(2) 90	43.4(3) 282	16.6(2) 307	8.7(2) 523	4.0(1) 705
Coverage	**97.0(1)** 1	**96.2(1)** 1	93.9(1) 2	70.9(2) 8	**63.9(2)** **33**	49.6(3) 191	25.7(3) 191	16.0(2) 437	**7.2(2)** 682
K-Median	**97.1(1)** 1	**96.3(1)** 1	**94.6(1)** 2	70.2(2) 9	63.4(2) 38	**50.0(2)** 214	23.5(2) 220	14.6(2) 434	6.1(2) 682
HV-Obs	**97.0(1)** 1	**96.3(1)** 1	94.4(1) 2	70.3(2) 9	61.6(2) 55	45.1(3) 260	23.5(3) 219	13.9(2) 484	4.4(1) 721
BC	96.5(1) 1	95.5(1) 1	92.3(1) 2	71.1(2) 7	62.9(2) 39	47.3(3) 198	27.9(3) 155	**16.1(2)** 386	6.7(2) **626**
CB	96.6(1) 1	95.7(1) 1	93.3(1) 2	**71.5(2)** **6**	63.0(2) 36	47.1(3) 221	**30.5(3)** **108**	**16.1(2)** **354**	4.9(1) **625**

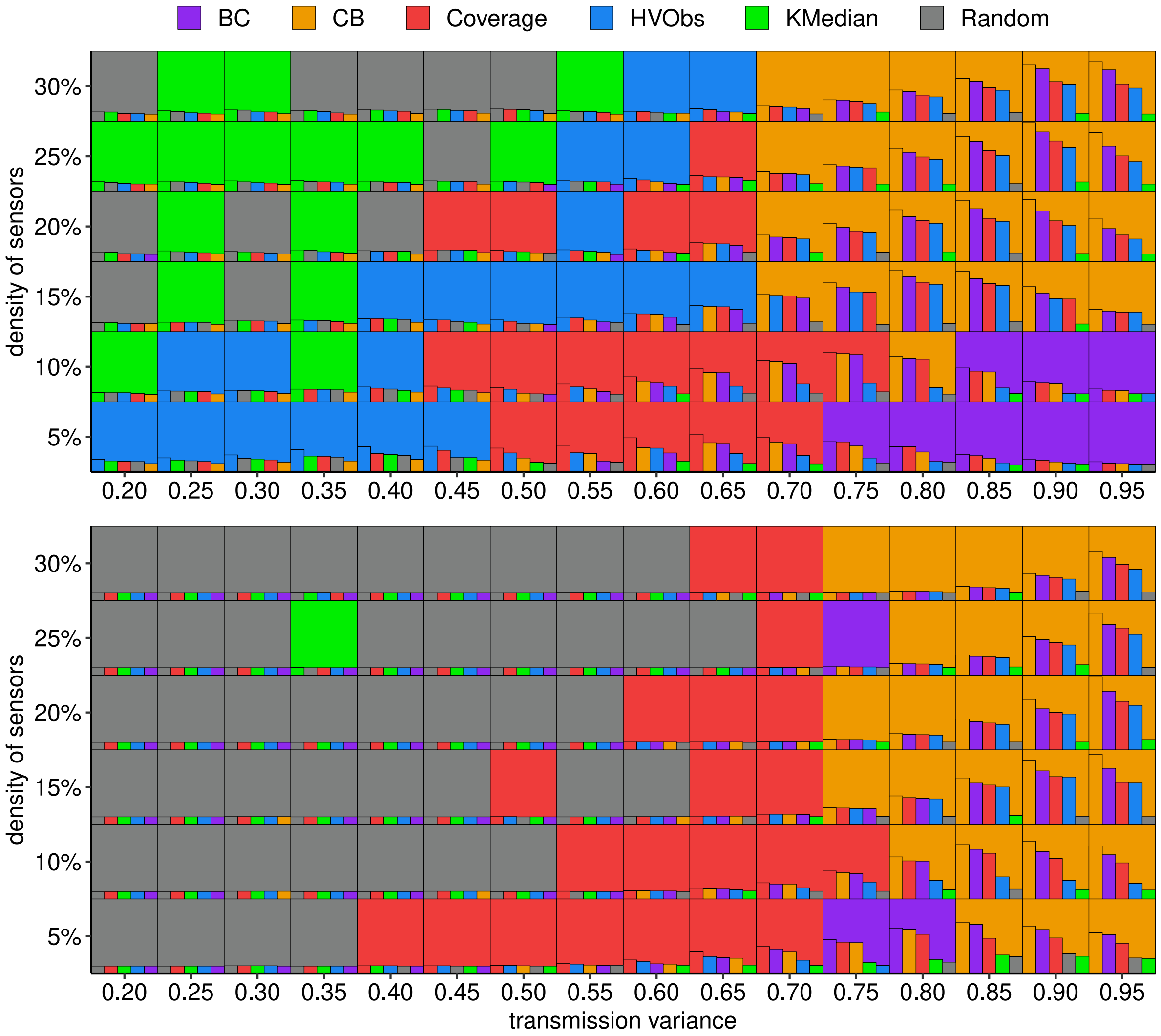

Fig. 13. Summary Diagrams for **Degree Sequence Algorithm** (n = 1000, $\langle k \rangle$ = 7.67). The highest bar for average precision (top diagram, $\xi = 0.9, \rho = 25\%$) represents difference of 19(1) percent points, while for 0.95-CSS the highest bar (bottom diagram, $\xi = 0.95, \rho = 20\%$) corresponds to 266 nodes. See Fig. 11 for detailed instruction how to read Summary Diagrams. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Summary Table for **Random Regular Graph** ($n = 1000$, $\langle k \rangle = 8$). See Table 4 for detailed instruction how to read Summary Table.

ξ →	⟨0.2; 0.5⟩			⟨0.5; 0.8⟩			⟨0.8; 0.95⟩		
ρ [%] →	20–30	10–20	5–10	20–30	10–20	5–10	20–30	10–20	5–10
Random	98.6(1) 1	97.7(1) 1	95.1(1) 1	73.1(2) 8	63.1(2) 48	48.0(2) 217	22.1(2) 216	12.0(2) 439	5.5(2) **657**
Coverage	**98.9(1)** 1	98.4(1) 1	96.5(1) 1	**76.9(2)** **4**	69.6(2) **17**	53.9(3) **151**	**32.1(3)** 122	19.7(2) 395	**8.3(2)** 675
K-Median	**98.8(1)** 1	**98.5(1)** 1	**97.4(1)** 1	75.6(2) 5	**70.0(2)** 18	**56.0(2)** 167	30.1(3) 144	**19.8(2)** **372**	7.6(2) 669
HV-Obs	**98.9(1)** 1	98.3(1) 1	96.3(1) 1	76.2(2) **4**	68.5(2) 21	51.7(3) 169	30.8(3) 136	19.3(2) 409	7.1(2) 671
BC	98.6(1) 1	97.6(1) 1	94.7(1) 2	73.2(2) 7	63.5(2) 43	48.1(3) 212	23.9(2) 213	13.5(2) 434	6.2(2) **657**
CB	98.7(1) 1	98.2(1) 1	96.7(1) 1	75.0(2) 5	66.8(2) 32	51.8(3) 212	31.7(3) **101**	16.8(2) 378	5.6(2) 666

transmission variance ξ influences the quality of source detection for various densities of sensors ρ. We investigate all networks as unweighted to limit the space of possible models' parameters. The heterogeneity of links is partially incorporated in the stochastic character of spreading process.

Figs. 3–10 show the results of numerical simulations in the most straightforward non-aggregated form. In each figure, the top two rows of charts present average precision, and the bottom two rows present Credible Set Size at the confidence level of 0.95 as a function of transmission variance ξ for six values of sensors' density (5%–30%). As evident, the transmission variance

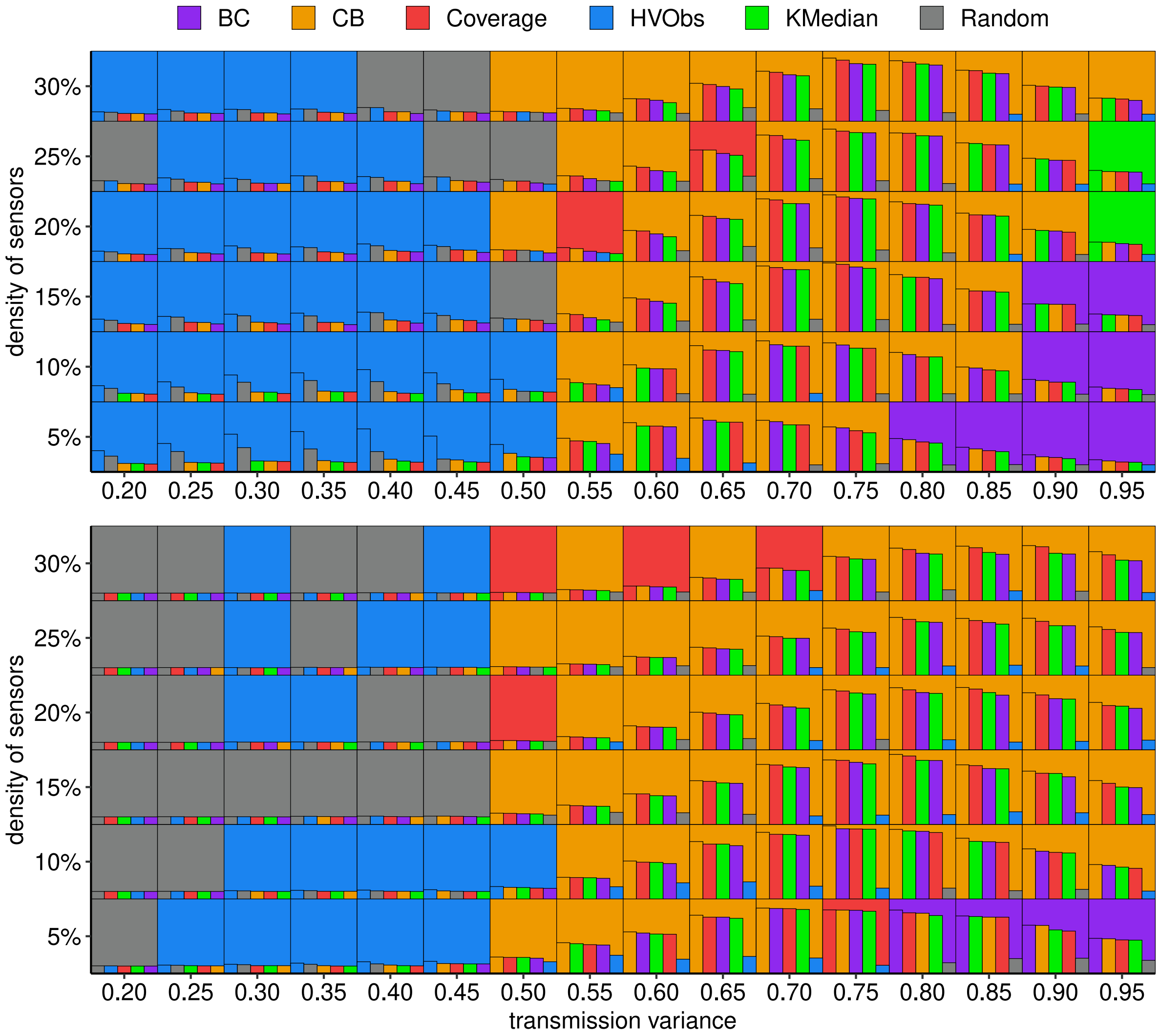

Fig. 14. Summary Diagrams for **Barabási–Albert** model ($n = 1000$, $\langle k \rangle = 7.98$). The highest bar for average precision (top diagram, $\xi = 0.75$, $\rho = 15\%$) represents difference of 22(1) percent points, while for 0.95-CSS the highest bar (bottom diagram, $\xi = 0.75$, $\rho = 10\%$) corresponds to 334 nodes. See Fig. 11 for detailed instruction how to read Summary Diagrams. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6
Summary Table for **Degree Sequence Algorithm** ($n = 1000$, $\langle k \rangle = 7.67$). See Table 4 for detailed instruction how to read Summary Table.

ξ →	⟨0.2; 0.5⟩			⟨0.5; 0.8⟩			⟨0.8; 0.95⟩		
ρ [%] →	20–30	10–20	5–10	20–30	10–20	5–10	20–30	10–20	5–10
Random	94.1(1) 2	93.1(1) 2	91.2(1) 2	67.7(2) 7	58.6(2) 29	45.4(3) 141	22.0(2) 180	12.1(2) 402	5.5(2) 623
Coverage	93.8(1) 2	93.2(1) 2	91.9(1) 2	70.7(2) 5	64.3(2) **10**	**52.3(3)** **75**	31.1(3) 73	20.1(2) 252	9.5(2) 545
K-Median	**94.2(1)** 2	93.1(1) 2	91.2(1) 2	67.9(2) 7	58.4(2) 29	45.2(3) 139	22.1(2) 180	12.1(2) 401	5.6(2) 619
HV-Obs	93.9(1) 2	**93.3(1)** 2	**92.4(1)** 2	70.5(2) 5	62.5(2) 15	47.9(3) 120	29.9(3) 86	18.1(2) 326	6.3(2) 634
BC	93.1(1) 2	92.1(1) 2	89.7(2) 2	70.4(2) 5	63.7(2) 11	50.3(2) 80	33.6(3) 59	21.4(2) 227	**9.9(2)** 495
CB	93.3(1) 2	92.5(1) 2	90.6(2) 2	**71.0(2)** 5	**64.5(2)** **10**	50.8(3) 85	**35.8(3)** **39**	**22.8(2)** **192**	9.2(2) **489**

is a major factor in the quality of source detection. For the low transmission variance, the precision provided by any reasonable algorithm for sensor placement is very high. In the opposite case, for the very high transmission variance, the quality of all methods is rather poor. The middle range of the variance transmission is characterized by the largest differences in precision and 0.95-CSS between the tested algorithms. The limit of this range depends on the density of sensors — it shifts towards the higher values of transmission variance with the higher density of sensors. This behavior can be observed for all types of synthetic and some real networks.

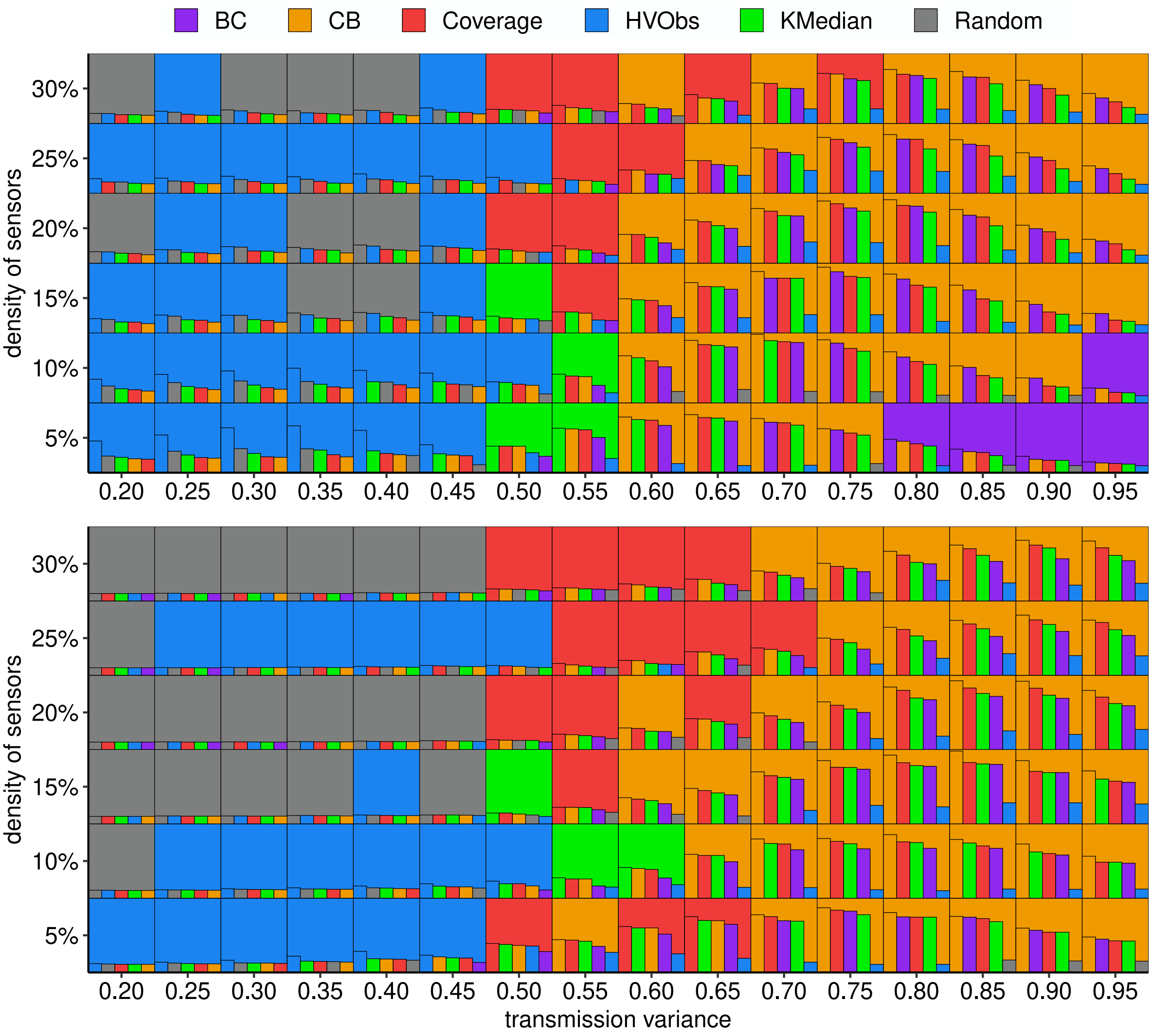

Fig. 15. Summary Diagrams for **Configuration Model** ($n = 1000$, $\langle k \rangle = 7.02$). The highest bar for average precision (top diagram, $\xi = 0.7, \rho = 10\%$) represents difference of 19(1) percent points, while for 0.95-CSS the highest bar (bottom diagram, $\xi = 0.85, \rho = 15\%$) corresponds to 273 nodes. See Fig. 11 for detailed instruction how to read Summary Diagrams. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
Summary Table for **Barabási–Albert** model ($n = 1000$, $\langle k \rangle = 7.98$). See Table 4 for detailed instruction how to read Summary Table.

ξ →	⟨0.2; 0.5⟩			⟨0.5; 0.8⟩			⟨0.8; 0.95⟩		
ρ [%] →	20–30	10–20	5–10	20–30	10–20	5–10	20–30	10–20	5–10
Random	89.7(1) 4	88.0(1) 6	83.1(2) 14	42.0(2) 264	35.2(2) 398	25.3(2) 540	3.3(1) 673	2.2(1) 745	1.2(1) 815
Coverage	88.6(1) 4	86.0(1) 7	80.2(2) 16	51.9(2) 97	46.4(2) 143	36.2(3) 248	15.1(2) 434	11.6(2) 523	7.0(2) 639
K-Median	87.7(1) 5	85.3(1) 8	80.4(2) 16	50.6(2) 105	45.5(2) 148	36.2(3) 248	15.0(2) 453	11.6(2) 528	6.7(2) 638
HV-Obs	**89.8(1)** 4	**89.3(1)** **4**	**87.5(2)** **6**	40.5(2) 276	34.7(2) 394	26.7(2) 531	3.2(1) 667	2.0(1) 749	1.1(1) 834
BC	88.1(1) 5	85.3(2) 8	79.5(2) 17	51.0(2) 107	45.8(2) 152	36.5(3) 248	14.7(2) 459	11.9(2) 529	**7.9(2)** 624
CB	88.6(1) 4	86.2(1) 7	80.8(2) 15	**52.2(2)** **93**	**47.2(2)** **136**	**37.7(3)** **238**	**15.6(2)** **422**	**12.3(2)** **508**	**7.7(2)** **623**

Summary Diagrams presented in Figs. 11–18 visualize extensive amount of complex results in an innovative way. In each figure, the top diagram refers to average precision, while the bottom diagram relates to 0.95-CSS. Each diagram consists of 96 tiles (16 values of ξ and 6 values of ρ). The color of the background in each tile indicates which algorithm provides the highest average precision (top) or the smallest 0.95-CSS (bottom) for the given pair (ξ, ρ). Moreover, inside each tile there are five colored bars, ordered from highest to lowest, that illustrate the ranking of the methods. The first bar relates to the best algorithm

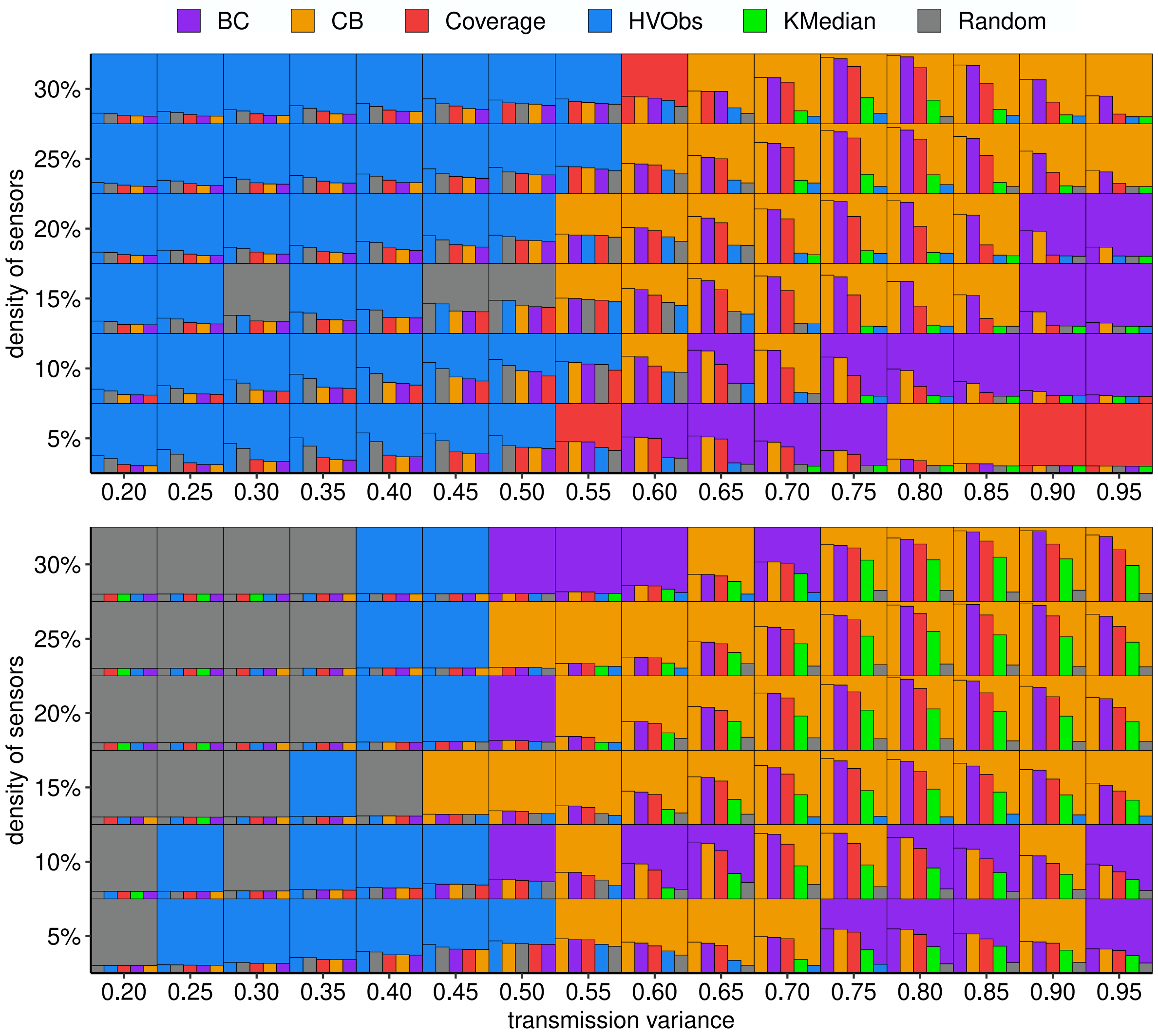

Fig. 16. Summary Diagrams for **University of California network** ($n = 1020$, $\langle k\rangle = 12.2$). The highest bar for average precision (top diagram, $\xi = 0.8$, $\rho = 30\%$) represents difference of 29(1) percent points, while for 0.95-CSS the highest bar (bottom diagram, $\xi = 0.9$, $\rho = 25\%$) corresponds to 489 nodes. See Fig. 11 for detailed instruction how to read Summary Diagrams. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 8
Summary Table for **Configuration Model** ($n = 1000$, $\langle k\rangle = 7.02$). See Table 4 for detailed instruction how to read Summary Table.

ξ →	⟨0.2; 0.5⟩			⟨0.5; 0.8⟩			⟨0.8; 0.95⟩		
ρ [%] →	20–30	10–20	5–10	20–30	10–20	5–10	20–30	10–20	5–10
Random	87.1(1) 6	85.5(1) 8	78.6(2) 32	40.9(2) 242	33.9(2) 355	23.7(2) 520	4.4(1) 638	2.5(1) 714	1.3(1) 800
Coverage	86.7(1) 6	84.7(1) 8	78.5(2) 26	49.3(2) 135	44.0(2) 181	34.5(3) 310	14.3(2) 434	9.8(2) 523	5.6(2) 653
K-Median	86.3(1) 7	85.0(1) 8	79.2(2) 25	47.7(2) 151	43.6(2) 189	34.4(3) 315	12.0(2) 459	8.8(1) 528	5.1(1) 654
HV-Obs	**87.6(1)** **5**	**86.5(1)** **7**	**83.3(2)** **14**	43.1(2) 230	35.4(2) 349	24.1(2) 511	6.3(1) 587	3.3(1) 694	1.2(1) 803
BC	85.2(1) 8	82.4(1) 13	75.4(2) 38	47.9(2) 162	43.0(2) 201	33.6(2) 333	15.0(2) 481	11.6(2) 535	**7.1(2)** 651
CB	86.0(1) 6	84.1(1) 9	78.1(2) 27	**49.4(2)** **132**	**45.1(2)** **169**	**35.7(2)** **304**	**16.4(2)** **413**	**12.6(2)** **483**	**7.0(2)** **630**

in the given tile, the second refers to second best and so on. The sixth algorithm in given tile is not shown, since it height would be zero. This is because the height of each bar shows the difference in average precision (or 0.95-CSS) between the method to which this bar refers and the worst (sixth) method in given tile. Then, the heights of bars from all tiles are scaled relative to the height of the highest bar among them, with a minimum height to recognize the color. This calibration allows to compare bars from various region of parameter space at cost of comparison of bars within one tile. Summary Diagrams give quick insight into results and

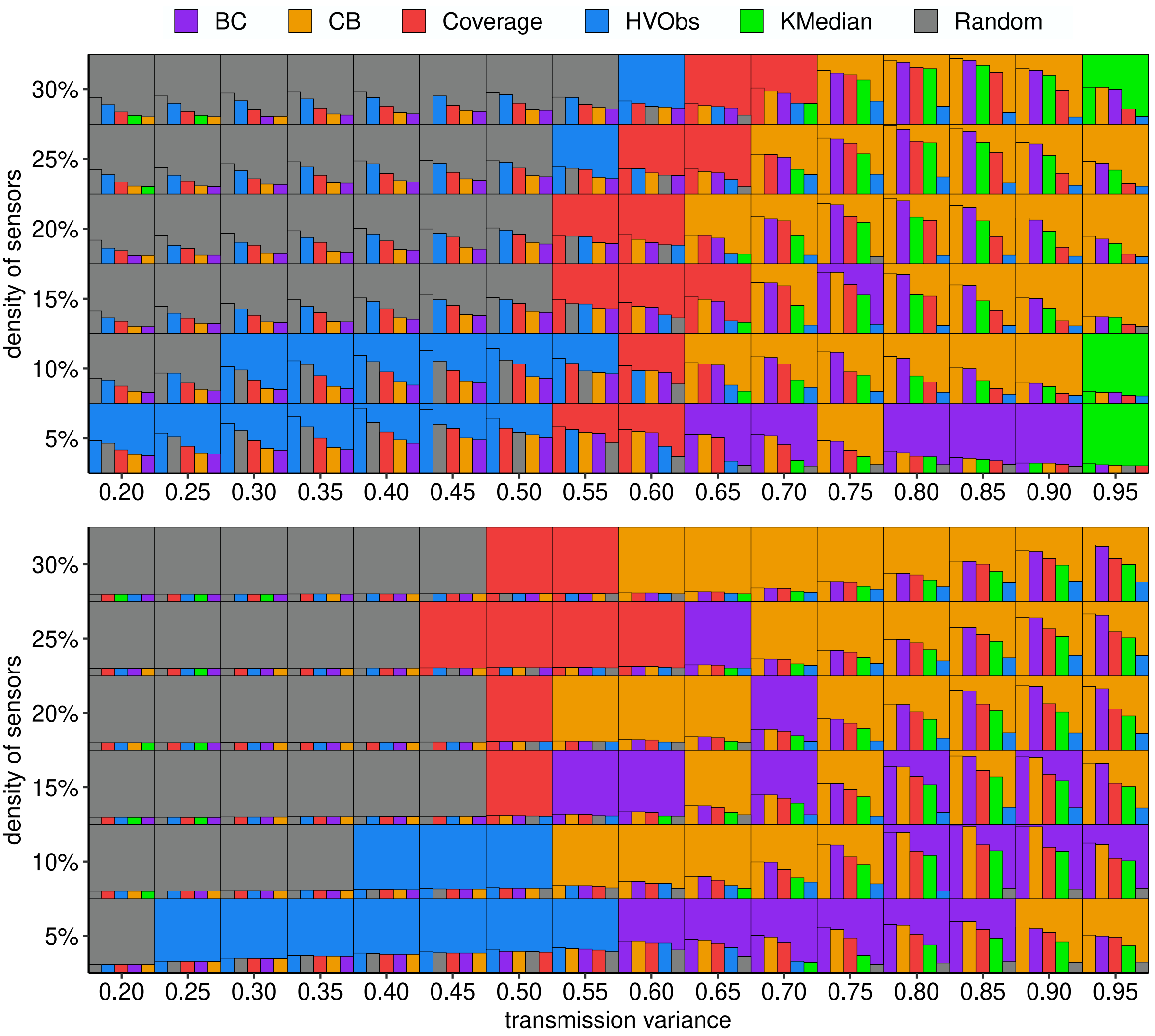

Fig. 17. Summary Diagrams for **University of Rovira i Virgili network** ($n = 1133$, $\langle k \rangle = 9.6$). The highest bar for average precision (top diagram, $\xi = 0.8, \rho = 25\%$) represents difference of 21(1) percent points, while for 0.95-CSS the highest bar (bottom diagram, $\xi = 0.85, \rho = 10\%$) corresponds to 535 nodes. See Fig. 11 for detailed instruction how to read Summary Diagrams. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 9
Summary Table for **University of California network** ($n = 1020$, $\langle k \rangle = 12$). See Table 4 for detailed instruction how to read Summary Table.

ξ →	⟨0.2; 0.5⟩			⟨0.5; 0.8⟩			⟨0.8; 0.95⟩		
ρ [%] →	20–30	10–20	5–10	20–30	10–20	5–10	20–30	10–20	5–10
Random	89.4(1) **3**	88.1(1) 4	84.6(2) 8	41.7(2) 359	36.5(2) 493	27.5(2) 612	2.1(1) 757	1.5(1) 818	0.9(1) 872
Coverage	87.9(2) **3**	85.1(2) 6	80.3(2) 17	52.5(2) 83	43.4(2) 178	32.6(2) 349	11.0(2) 419	4.9(1) 567	2.2(1) 713
K-Median	84.8(2) 6	81.1(2) 14	75.9(2) 58	39.9(2) 188	30.7(2) 337	22.3(2) 511	3.7(1) 557	1.8(1) 676	1.0(1) 775
HV-Obs	**90.4(1)** **3**	**89.1(1)** **3**	**87.3(2)** **5**	43.2(2) 389	36.9(2) 526	28.2(2) 626	2.5(1) 769	1.7(1) 818	0.8(1) 887
BC	87.0(2) 4	85.0(1) 5	80.2(2) 15	54.7(2) 59	48.4(2) 125	**36.0(2)** 306	19.6(3) 329	**12.2(2)** 499	**4.4(1)** 678
CB	87.2(2) 4	85.3(1) 5	80.5(2) 15	**55.3(2)** **55**	**48.7(2)** **122**	**36.0(2)** **305**	**20.2(3)** **318**	**12.1(2)** **498**	4.1(1) **673**

show in legible way which algorithms lead in which regions of the parameter space.

To better understand and interpret these complex results, we partition the parameter space into nine regions of similar size and shared boundaries which correspond to low, medium and high transmission variance and density of sensors. Tables 4–11 present average values for precision (top numbers in cells, in percentages) and Credible Set Size (bottom numbers in cells) in these regions.

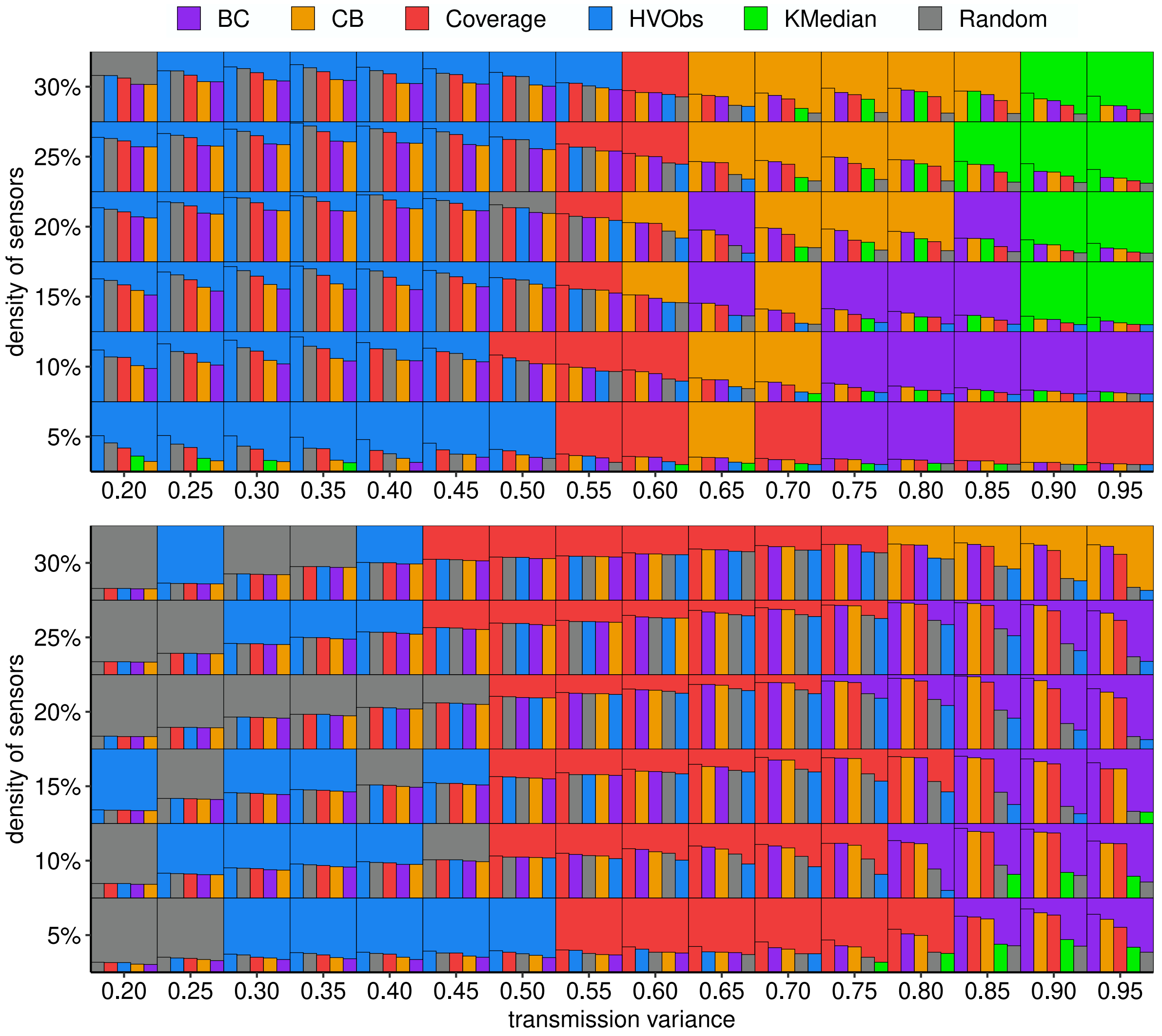

Fig. 18. Summary Diagrams for **Infectious network** ($n = 410$, $\langle k\rangle = 13.5$). The highest bar for average precision (top diagram, $\xi = 0.35, \rho = 25\%$) represents difference of 33(1) percent points, while for 0.95-CSS the highest bar (bottom diagram, $\xi = 0.85, \rho = 20\%$) corresponds to 173 nodes. See Fig. 11 for detailed instruction how to read Summary Diagrams. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In each table, the first three numerical columns from the left refer to low, three in the middle to medium, and the last three to high transmission variance ξ. Similarly, columns $\{1, 4, 7\}$ correspond to high, $\{2, 5, 8\}$ to middle, and $\{3, 6, 9\}$ to low density of sensors ρ. This arrangement of columns causes the precision decreases from left to right side of the table. The best results in each region are printed in bold. The uncertainty of precision is given by the confidence interval at the level 0.95.

6.1. Tests on synthetic networks

For each model of network and each value of ρ we execute the following script 100 times:

1. Generate a new graph.
2. Find sets of sensors according to six different selection strategies (Random, Coverage, K-Median, HV-Obs, BC and CB).
3. For each value of ξ repeat 100 times:
   (a) Simulate spread from random source using SI model.
   (b) Locate the source six times using different sets of sensors.

As a result, the values of average precision and the average Credible Set Size at the confidence level 0.95 for each point (ρ, ξ) are computed from 10^4 attempts to locate the source. Table 2 presents the characteristics of studied networks.

6.1.1. Erdős-Rényi Graph (ER)

This random graph is constructed by connecting every pair of nodes with probability p. The resulting network has binomial degree distribution with the average degree $\langle k\rangle = pn$. The average path length (APL) scales linearly with $\ln n$ (it is a property of so-called *small-world networks*) and it is much smaller than the number of edges. The global clustering coefficient is p since the probability that two connected nodes have a common neighbor is uniform for all nodes and equal to p. According to Table 4, Coverage and K-Median are the best approaches for sensor placement in ER graph for the moderate transmission variance range $\xi \in \langle 0.5; 0.8\rangle$ and the density of sensors below or equal to 20%. For larger budgets $\rho \in \langle 20\%; 30\%\rangle$ Collective Betweenness delivers the highest quality of source detection in that range of transmission variance. Also CB is the very effective when spread is highly stochastic $\xi \in \langle 0.8; 0.95\rangle$ and the density of sensors higher or equal to 10%. For the low transmission variance range all methods give the same values for the Credible Set Sizes with

Table 10
Summary Table for **University of Rovira i Virgili network** ($n = 1133$, $\langle k \rangle = 9.6$). See detailed instruction how to read Summary Table under Table 4.

ξ →	$\langle 0.2; 0.5 \rangle$			$\langle 0.5; 0.8 \rangle$			$\langle 0.8; 0.95 \rangle$		
ρ [%] →	20–30	10–20	5–10	20–30	10–20	5–10	20–30	10–20	5–10
Random	**77.5(1)** **4**	**75.2(2)** 5	71.6(2) 7	39.1(2) 115	32.6(2) 232	24.3(2) 399	4.0(1) 566	2.5(1) 738	1.5(1) 880
Coverage	73.2(2) **4**	71.7(2) **5**	68.9(2) 8	45.7(2) 32	39.6(2) 76	30.4(2) 208	12.4(2) 255	6.9(1) 419	3.3(1) 630
K-Median	69.2(2) 7	66.0(2) 12	60.3(2) 57	41.1(2) 58	33.6(2) 119	22.9(2) 313	16.1(2) 299	9.9(2) 449	4.6(1) 670
HV-Obs	75.5(1) **4**	74.3(2) **4**	**74.1(2)** **4**	40.8(2) 99	33.9(2) 208	26.9(2) 371	5.0(1) 480	2.9(1) 707	1.7(1) 899
BC	70.6(2) 5	68.5(2) 6	65.4(2) 9	45.2(2) 23	40.1(2) 40	31.1(2) 138	19.1(2) **137**	13.5(2) **251**	6.2(1) **545**
CB	70.8(2) 5	68.8(2) 6	66.1(2) 9	**45.9(2)** **22**	**40.7(2)** **39**	**31.4(2)** 143	**19.9(2)** **128**	**14.0(2)** **252**	**6.3(2)** 549

the confidence level 0.95, but K-Median has the highest average precision.

6.1.2. Random regular graph (RRG)

Nodes in RRG are connected at random, but with the constraint that each node has the same degree. RRG has a slightly lower global clustering coefficient and a higher average path length than ER graph with the same number of nodes and edges. Table 5 shows that K-Median and Coverage provides the highest quality of source detection, but for the most stochastic processes and largest budgets, Collective Betweenness gives the smallest Credible Set Sizes at the confidence level 0.95 and it is the second best in the average precision. In this region (the highest transmission variance and density of sensors), difference between the efficiency of Collective Betweenness and Betweenness Centrality is the largest among all studied networks and all regions. In fact, for $\xi = 0.95$ and $\rho \in \langle 25\%, 30\% \rangle$ CB provides the highest average precision, while BC is the second worse in that region (see Fig. 12).

6.1.3. Degree sequence algorithm (DSA)

This model is a random synthetic network with the high clustering coefficient. The algorithm takes a sequence of nodes' degrees as an input and global clustering coefficient as parameter, and returns the list of links forming the graph. The degrees of nodes are selected from the Poisson distribution for easier comparison with ER graph. The results presented in Table 6 indicate a significant advantage of Collective Betweenness over the other methods for the transmission variance $\xi \geqslant 0.5$ and the density of sensors $\rho \geqslant 10\%$. For the smaller budgets $\rho \leqslant 10\%$, Coverage and Betweenness Centrality provide higher precision than CB. In the case of low transmission variance $\xi \leqslant 0.5$, HV-Obs gives the highest precision for $\rho \geqslant 20\%$. In contrast to the results for ER and RRG, here K-Median performs poorly, with exception of cases with the lowest values of transmission variance and the largest numbers of sensors.

6.1.4. Barabási–Albert Model (BA)

This algorithm uses the preferential attachment rule to generate a scale-free network, which has a smaller average path length and a higher clustering coefficient than Erdős–Rényi graph (but sill much smaller than for real social networks). Fig. 14 reveals fragmentation of parameter space into three distinct regions. The first region includes all the densities of sensors with the transmission variance $\xi < 0.5$. In this region, HV-Obs provides the highest quality of source localization (the lower density of sensors is, the bigger is the advantage of HV-Obs over competitors). The second region, in which CB is the most effective method, includes all the densities of sensors with the transmission variance $\xi > 0.5$, except of the corner with the highest transmission variance and the lowest density of sensors, where BC is the best method.

6.1.5. Configuration model (CM)

This model generates a random graph from a given degree sequence. In our studies, we use the power law degree distribution for easier comparison with the Barabási–Albert model. Although the results shown in Table 8 are consistent with the results for BA network, Fig. 15 reveals a small subregion (with the medium transmission variance $\sigma = 0.55$ and the high density of sensors $\rho \geqslant 15\%$) where the leading method is Coverage.

6.1.6. Summary of tests on synthetic networks

The results in Tables 4–8 show that for the low transmission variance $\xi \in \langle 0.2; 0.5 \rangle$, the profit from choosing the particular set of sensors is very small when the density of sensors exceed 10%. In the case of lower density of sensors, the highest quality of source detection is provided by HV-Obs (for the Degree Sequence Algorithm, Configuration Model, Barabási–Albert model) and K-Median (for the Erdős–Rényi graph, Random Regular Graph). On the opposite side, for the high transmission variance $\xi \in \langle 0.8; 0.95 \rangle$, the gain from using specific sensors is much higher, but the choice of algorithm depends on the network type. For Erdős–Rényi the best are Coverage (when $\rho \leqslant 10\%$) and Collective Betweenness (when $\rho > 10\%$). Coverage is also the best choice for Random Regular Graph. For the Degree Sequence Algorithm, Barabási–Albert and Configuration Model, the highest precision and the smallest Credible Set Size at the confidence level 0.95 are provided by the Collective Betweenness. Also in the middle range of transmission variance $\xi \in \langle 0.5; 0.8 \rangle$ The Collective Betweenness is the best algorithm for these graphs (except DSA with low density of sensors), but for ER and RRG it gives way to Coverage and K-Median.

6.2. Tests on real networks

The real networks used in the study come from the Koblenz Network Collection [49] (KONTECT). Table 3 presents the characteristics of these networks. For each real network and each pair of (ρ, ξ) we simulate the spread from a random source and locate the source using different sets of sensors 10^4 times.

6.2.1. University of California

This network contains information about the message exchanges between the users of an online community of students from the University of California, Irvine [50] A node represents a user and the directed multiple edges represent messages. We transform the network to undirected one in the same way as Spinelli et al. [32], by aggregating all edges between a given pair of nodes and leaving only the connections with at least one edge in both directions. Then we remove iteratively all nodes with

Table 11
Summary Table for **Infectious network** ($n = 410$, $\langle k \rangle = 13.5$). See Table 4 for detailed instruction how to read Summary Table.

ξ →	$\langle 0.2; 0.5\rangle$			$\langle 0.5; 0.8\rangle$			$\langle 0.8; 0.95\rangle$		
ρ [%] →	20–30	10–20	5–10	20–30	10–20	5–10	20–30	10–20	5–10
Random	78.8(2) **6**	73.2(2) **9**	59.7(3) 19	31.6(2) 51	24.1(2) 76	14.9(2) 107	4.0(1) 181	2.8(1) 234	1.9(1) 272
Coverage	77.1(2) **6**	71.7(2) 10	59.8(3) 20	36.5(2) **30**	28.1(2) **44**	**18.8(2)** **70**	8.2(2) 100	4.5(1) 136	3.1(1) 184
K-Median	52.0(2) 84	46.6(2) 84	44.9(3) 68	25.4(2) 167	17.2(2) 178	12.0(2) 155	13.1(2) 228	7.1(1) 256	2.8(1) 265
HV-Obs	**79.8(2)** **6**	**74.9(2)** **9**	**64.3(2)** **17**	30.3(2) 56	23.6(2) 94	16.1(2) 120	2.8(1) 202	2.5(1) 271	1.9(1) 311
BC	72.7(2) 9	66.4(2) 13	53.2(2) 27	36.8(2) 31	28.5(2) 46	18.2(2) 81	11.4(2) **77**	**7.3(1)** **122**	**4.1(1)** **177**
CB	72.6(2) 9	67.4(2) 13	54.7(3) 24	**37.2(2)** 32	**28.8(2)** 48	18.4(2) 82	**11.7(2)** 79	6.8(1) 130	3.7(1) 185

less than two connections until the minimum node degree in the network is two. The results of numerical tests contained in Table 9 are very similar to the results for Barabási–Albert model (Table 7). The most effective method for the low transmission variance ξ is HV-Obs, while for the medium and high ξ, the best quality of source detection is provided by Betweenness Centrality and Collective Betweenness.

6.2.2. University of rovira i virgili

This is the email communication network at the University Rovira i Virgili in Tarragona in the south of Catalonia in Spain [51]. An undirected link between two nodes (users) is created when at least one email was sent from one user to another. The results obtained for this network are in line with the results for synthetic scale-free networks and the University of California network. The main difference is observed for the low transmission variance $\xi \in \langle 0.2; 0.5\rangle$ and $\rho \geqslant 10\%$, where all sets of sensors perform worse than random ones (see Table 10).

6.2.3. Infectious

This network contains human face-to-face interactions during the exhibition INFECTIOUS: STAY AWAY in 2009 at the Science Gallery in Dublin [52]. Edges represent contacts which lasted for at least 20 s. Only the data from the day with the most interactions was used. The network is characterized by highest average degree and clustering coefficient among all tested graphs. Table 11 shows that the low transmission variance region is again dominated by HV-Obs, in particular for $\rho \leqslant 10\%$. The best option for the high transmission variance is Betweenness Centrality, and Collective Betweenness is the first choice for medium range of ξ, except of a region with the low density of sensors $\rho \leqslant 10\%$, where Coverage performs better.

6.2.4. Summary of tests on real networks

In our tests, locating the source of information is much more challenging on real networks than on artificial ones due to high clustering coefficient of such networks. On the other hand, employing the right strategy of sensor placement can be more fruitful in case of the real systems than for the synthetic ones. The common denominator for all the real networks studied in this paper is the advantage of betweenness-based algorithms over the rest of methods for the medium and high transmission variance. In the case of low transmission variance, best strategy for sensor placement are HV-Obs and Random (the latter only for the University of Rovira i Virgili network). For the Infectious network also Coverage performs well for medium transmission variance (in particular for $\xi = 0.55$).

7. Discussion

In this article we review the methods of sensor placement for the source localization in complex networks, both real and synthetic, using a well established propagation model – Susceptible–Infected – and the Pinto–Thiran–Vetterli localization algorithm. We have selected four vastly acknowledged methods for sensor placement – Betweenness Centrality (BC), High Coverage Rate (Coverage), K-Median, High Variance Observers (HV-Obs) – and also have introduced our own method called Collective Betweenness (CB). We compare all of these methods with each other and with a baseline method — random selection. As our main evaluation metrics, we use an average precision of identifying the actual source of the spread and introduce a new metric called Credible Set Size that we believe to be more useful in possible real world scenarios as it conveys a very practical notion of "how many nodes do I need to check to say the source is one of them with credibility α?". The study was conducted over a large range of values for the two main parameters affecting the propagation and source identification — the transmission variance ξ and the density of sensors ρ.

As shown in Tables 4–11 and in Figs. 3–10, right choice of sensor placement can significantly increase precision and reduce Credible Set Size. However, the gain of doing so varies from very high values in some cases to moderate in others. For example, for the high values of ξ and ρ (see third column in Tables 4–11) the increase of average precision can reach even 18 percent (from 2.1 to 20.2, for the University of California network) in comparison to the baseline random method. On the other hand, for low values of transmission variance $\xi \in \langle 0.2; 0.5\rangle$, the performance of different sets of sensors is very similar and a noticeable gain can be observed only for the scale-free networks when the density of sensors is rather low. Figs. 11–18 show which methods outperforms the others for the given transmission variance ξ and density of sensors ρ. These intricate, colorful mosaics can be difficult to interpret but some patterns recur across different networks. The first thing that catches the eye is the difference between networks with a narrow degree distribution (Erdős–Rényi, Random Regular Graph and Degree Sequence Algorithm) and scale-free networks. Unlike the former, performance of the later splits the parameter space into two clear parts. The summary maps for Barabási–Albert, Configuration Model, University of California, University of Rovira and Infectious networks are visibly divided into two parts. The left part corresponds to the low transmission variance, and it is dominated by HV-Obs and Random. The right part represents spreading with higher stochasticity. It is more favorable to CB or BC (with the exception of the Infectious network, where K-Median also appears as a leader in a small region). However, in case of ER, RRG and DSA networks, Coverage and K-Median are also performing quite well along with HV-Obs and CB and the domains of the particular methods are usually more fragmented for these networks than for the real and scale-free networks.

In summary, among all the studied algorithms for sensor placement, two of them perform particularly well. The first one is High Variance Observers which outperforms the others when transmission variance is low, and the second one is Collective Betweenness which provides the highest quality of source localization when spreading is highly stochastic and unpredictable. Despite the large number of tests carried out in this review, there is still place for additional research on this topic. This work has been done for unweighted networks, while the connections in real-world networks are usually not identical and thus future testing of the source localization on a weighted graphs would be of great value. Similarly, in our study, we perform source location when the signal has already reached all sensors whereas one could imagine situations when we are trying to locate the source as soon as a small fraction of sensors is reached by the spread and under such conditions one could imagine significantly different results than presented here by us.

CRediT authorship contribution statement

Robert Paluch: Conceptualization, Methodology, Software, Investigation, Writing - original draft. **Łukasz G. Gajewski:** Methodology, Writing - original draft, Writing - review & editing. **Janusz A. Hołyst:** Supervision, Writing - review & editing. **Boleslaw K. Szymanski:** Resources, Supervision, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work was partially supported as RENOIR Project by the European Union Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No 691152 and by Ministry of Science and Higher Education (Poland), Grant Nos. 34/H2020/2016, 329025/PnH/2016 and the National Science Centre, Poland Grant No. 2015/19/B/ST6/02612. R.P. was partially supported by the National Science Centre, Poland, agreement No UMO-2019/32/T/ST6/00173, and by PLGrid Infrastructure. J.A.H. was partially supported by the Russian Science Foundation, Agreement #17-71-30029 with co-financing of Bank Saint Petersburg, Russia. B.K.S. was partially supported by the U.S. Army Research Office Grant No. W911NF-16-1-0524, U.S. Office of Naval Research Grant No. N00014-15-1-2640 and by the National Science Centre, Poland Grant No. 2016/21/B/ST6/01463.

References

[1] D. Westerman, P.R. Spence, B. Van Der Heide, Social media as information source: Recency of updates and credibility of information, J. Comput.-Mediated Commun. 19 (2) (2014) 171–183.

[2] M. Smith, G. Mulrain, Equi-failure: The national security implications of the equifax hack and a critical proposal for reform, J. Nat'l Sec. L. & Pol'y 9 (2017) 549.

[3] C. Cadwalladr, E. Graham-Harrison, Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach, Guardian 17 (2018) 22.

[4] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (6380) (2018) 1146, http://dx.doi.org/10.1126/science.aap9559, http://science.sciencemag.org/content/359/6380/1146.abstract.

[5] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H.E. Stanley, W. Quattrociocchi, The spreading of misinformation online, Proc. Natl. Acad. Sci. 113 (3) (2016) 554, http://dx.doi.org/10.1073/pnas.1517441113, http://www.pnas.org/content/113/3/554.abstract.

[6] P. Törnberg, Echo chambers and viral misinformation: Modeling fake news as complex contagion, PLOS ONE 13 (9) (2018) 1–21, http://dx.doi.org/10.1371/journal.pone.0203958.

[7] R.K. Garrett, Echo chambers online?: Politically motivated selective exposure among internet news users1, J. Comput.-Mediated Commun. 14 (2) (2009) 265–285, https://doi.org/10.1111/j.1083-6101.2009.01440.x.

[8] C. Shao, G.L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, F. Menczer, The spread of low-credibility content by social bots, Nature Commun. 9 (1) (2018) 4787, https://doi.org/10.1038/s41467-018-06930-7.

[9] C. Shao, P.-M. Hui, L. Wang, X. Jiang, A. Flammini, F. Menczer, G.L. Ciampaglia, Anatomy of an online misinformation network, PLOS ONE 13 (4) (2018) 1–23, http://dx.doi.org/10.1371/journal.pone.0196087.

[10] C. Fraser, C.A. Donnelly, S. Cauchemez, W.P. Hanage, M.D. Van Kerkhove, T.D. Hollingsworth, J. Griffin, R.F. Baggaley, H.E. Jenkins, E.J. Lyons, T. Jombart, W.R. Hinsley, N.C. Grassly, F. Balloux, A.C. Ghani, N.M. Ferguson, A. Rambaut, O.G. Pybus, H. Lopez-Gatell, C.M. Alpuche-Aranda, I.B. Chapela, E.P. Zavala, D.M.E. Guevara, F. Checchi, E. Garcia, S. Hugonnet, C. Roth, Pandemic potential of a strain of influenza a (H1N1): Early findings, Science 324 (5934) (2009) 1557, http://dx.doi.org/10.1126/science.1176062, http://science.sciencemag.org/content/324/5934/1557.abstract.

[11] G. Neumann, T. Noda, Y. Kawaoka, Emergence and pandemic potential of swine-origin H1N1 influenza virus, Nature 459 (7249) (2009) 931–939, http://dx.doi.org/10.1038/nature08157.

[12] D.S. Hui, E. I Azhar, T.A. Madani, F. Ntoumi, R. Kock, O. Dar, G. Ippolito, T.D. Mchugh, Z.A. Memish, C. Drosten, et al., The continuing 2019-ncov epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China, Int. J. Infect. Dis. 91 (2020) 264–266.

[13] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Phys. Rev. Lett. 86 (2001) 3200–3203, http://dx.doi.org/10.1103/PhysRevLett.86.3200, https://link.aps.org/doi/10.1103/PhysRevLett.86.3200.

[14] X. Fan, Y. Xiang, Modeling the propagation of peer-to-peer worms, Future Gener. Comput. Syst. 26 (8) (2010) 1433–1443, http://dx.doi.org/10.1016/j.future.2010.04.009, http://www.sciencedirect.com/science/article/pii/S0167739X10000737.

[15] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, A. Vespignani, Epidemic processes in complex networks, Rev. Modern Phys. 87 (2015) 925–979, http://dx.doi.org/10.1103/RevModPhys.87.925, https://link.aps.org/doi/10.1103/RevModPhys.87.925.

[16] Z. Masood, R. Samar, M.A.Z. Raja, Design of fractional order epidemic model for future generation tiny hardware implants, Future Gener. Comput. Syst. 106 (2020) 43–54, http://dx.doi.org/10.1016/j.future.2019.12.053, http://www.sciencedirect.com/science/article/pii/S0167739X19317170.

[17] S.D. Bhattacharjee, W.J. Tolone, V.S. Paranjape, Identifying malicious social media contents using multi-view context-aware active learning, Future Gener. Comput. Syst. 100 (2019) 365–379, http://dx.doi.org/10.1016/j.future.2019.03.015, http://www.sciencedirect.com/science/article/pii/S0167739X18307349.

[18] Z. Tan, D. Wu, T. Gao, I. You, V. Sharma, AIM: Activation increment minimization strategy for preventing bad information diffusion in OSNs, Future Gener. Comput. Syst. 94 (2019) 293–301, http://dx.doi.org/10.1016/j.future.2018.11.038, http://www.sciencedirect.com/science/article/pii/S0167739X18319058.

[19] D. Brockmann, D. Helbing, The hidden geometry of complex, network-driven contagion phenomena, Science 342 (6164) (2013) 1337–1342, http://dx.doi.org/10.1126/science.1245200, http://www.sciencemag.org/cgi/doi/10.1126/science.1245200.

[20] M.E.J. Newman, Spread of epidemic disease on networks, Phys. Rev. E 66 (2002) 016128, http://dx.doi.org/10.1103/PhysRevE.66.016128, https://link.aps.org/doi/10.1103/PhysRevE.66.016128.

[21] A.Y. Lokhov, M. Mézard, H. Ohta, L. Zdeborová, Inferring the origin of an epidemic with a dynamic message-passing algorithm, Phys. Rev. E 90 (1) (2014) 012801, http://dx.doi.org/10.1103/PhysRevE.90.012801, https://link.aps.org/doi/10.1103/PhysRevE.90.012801.

[22] Z. Shen, S. Cao, W.-X. Wang, Z. Di, H.E. Stanley, Locating the source of diffusion in complex networks by time-reversal backward spreading, Phys. Rev. E 93 (3) (2016) 032301, http://dx.doi.org/10.1103/PhysRevE.93.032301, https://link.aps.org/doi/10.1103/PhysRevE.93.032301.

[23] H. Wang, An universal algorithm for source location in complex networks, Physica A 514 (2019) 620–630, http://dx.doi.org/10.1016/j.physa.2018.09.114, http://www.sciencedirect.com/science/article/pii/S0378437118312470.

[24] S. Xu, C. Teng, Y. Zhou, J. Peng, Y. Zhang, Z.K. Zhang, Identifying the diffusion source in complex networks with limited observers, Physica A 527 (2019) 121267, http://dx.doi.org/10.1016/j.physa.2019.121267, https://doi.org/10.1016/j.physa.2019.121267.

[25] P.C. Pinto, P. Thiran, M. Vetterli, Locating the source of diffusion in large-scale networks, Phys. Rev. Lett. 109 (6) (2012) 1–5, http://dx.doi.org/10.1103/PhysRevLett.109.068702, arXiv:1208.2534.

[26] X. Li, X. Wang, C. Zhao, X. Zhang, D. Yi, Optimal identification of multiple diffusion sources in complex networks with partial observations, in: Y. Liu, L. Wang, L. Zhao, Z. Yu (Eds.), Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery, Springer International Publishing, Cham, ISBN: 978-3-030-32456-8, 2020, pp. 214–223.

[27] Ł.G. Gajewski, K. Suchecki, J.A. Hołyst, Multiple propagation paths enhance locating the source of diffusion in complex networks, Physica A 519 (2019) 34–41, http://dx.doi.org/10.1016/j.physa.2018.12.012, http://www.sciencedirect.com/science/article/pii/S0378437118315176.

[28] R. Paluch, X. Lu, K. Suchecki, B.K. Szymański, J.A. Hołyst, Fast and accurate detection of spread source in large complex networks, Sci. Rep. (ISSN: 2045-2322) 8 (1) (2018) 2508, http://dx.doi.org/10.1038/s41598-018-20546-3, https://doi.org/10.1038/s41598-018-20546-3.

[29] X. Li, X. Wang, C. Zhao, X. Zhang, D. Yi, Locating the source of diffusion in complex networks via Gaussian-based localization and deduction, Appl. Sci. 9 (18) (2019) http://dx.doi.org/10.3390/app9183758, https://www.mdpi.com/2076-3417/9/18/3758.

[30] X. Zhang, Y. Zhang, T. Lv, Y. Yin, Identification of efficient observers for locating spreading source in complex networks, Physica A 442 (2016) 100–109, http://dx.doi.org/10.1016/j.physa.2015.09.017, http://dx.doi.org/10.1016/j.physa.2015.09.017.

[31] U. Brandes, A faster algorithm for betweenness centrality, J. Math. Sociol. 25 (2001) 163–177.

[32] B. Spinelli, L.E. Celis, P. Thiran, Observer placement for source localization: The effect of budgets and transmission variance, 54th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2016 (2017) 743–751, http://dx.doi.org/10.1109/ALLERTON.2016.7852307, arXiv:1608.04567.

[33] O. Kariv, S.L. Hakimi, An algorithmic approach to network location problems. II: The p-medians, SIAM J. Appl. Math. 37 (3) (1979) 539–560.

[34] L.E. Celis, F. Pavetić, B. Spinelli, P. Thiran, Budgeted sensor placement for source localization on trees, Electron. Notes Discrete Math. 50 (2015) 65–70, http://dx.doi.org/10.1016/j.endm.2015.07.012, http://www.sciencedirect.com/science/article/pii/S1571065315001675.

[35] B. Spinelli, L. Celis, P. Thiran, A general framework for sensor placement in source localization, IEEE Trans. Netw. Sci. Eng. 6 (2) (2019) 86–102, http://dx.doi.org/10.1109/TNSE.2017.2787551.

[36] B. Spinelli, L.E. Celis, P. Thiran, Back to the source: An online approach for sensor placement and source localization, in: Proceedings of the 26th International Conference on World Wide Web, in: WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2017, pp. 1151–1160, http://dx.doi.org/10.1145/3038912.3052584.

[37] Y. Zhang, X. Zhang, B. Zhang, An observer deployment algorithm for locating the diffusion source timely in social network, in: 2016 2nd Workshop on Advanced Research and Technology in Industry Applications (WARTIA-16), Atlantis Press, 2016/05, http://dx.doi.org/10.2991/wartia-16.2016.333.

[38] S. Zejnilović, J. Gomes, B. Sinopoli, Sequential observer selection for source localization, in: 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), IEEE, 2015, pp. 1220–1224.

[39] S. Zejnilović, J. Xavier, J. Gomes, B. Sinopoli, Selecting observers for source localization via error exponents, in: 2015 IEEE International Symposium on Information Theory (ISIT), IEEE, 2015, pp. 2914–2918.

[40] S. Zejnilović, J. Gomes, B. Sinopoli, Network observability and localization of the source of diffusion based on a subset of nodes, in: 2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2013, pp. 847–852.

[41] M. Fang, P. Shi, W. Shang, X. Yu, T. Wu, Y. Liu, Locating the source of asynchronous diffusion process in online social networks, IEEE Access 6 (2018) 17699–17710, http://dx.doi.org/10.1109/ACCESS.2018.2817553.

[42] C. Shi, Q. Zhang, T. Chu, Observer selection for source identification on complex networks, in: 2019 Chinese Control Conference (CCC), Technical Committee on Control Theory, Chinese Association of Automation, 2019, pp. 7996–8000.

[43] N.T. Bailey, et al., The Mathematical Theory of Infectious Diseases and Its Applications, Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975.

[44] J. Berry, W.E. Hart, C.A. Phillips, J.G. Uber, J.-P. Watson, Sensor placement in municipal water networks with temporal integer programming models, J. Water Resour. Plan. Manag. 9496 (2006) http://dx.doi.org/10.1061/(ASCE)0733-9496(2006)132.

[45] M. Vijaymeena, K. Kavitha, A survey on similarity measures in text mining, Mach. Learn. Appl. 3 (2) (2016) 19–28.

[46] P. Erdős, A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci. 5 (1) (1960) 17–60.

[47] L.S. Heath, N. Parikh, Generating random graphs with tunable clustering coefficients, Physica A (2011) http://dx.doi.org/10.1016/j.physa.2011.06.052.

[48] R. Albert, A.L. Barabasi, Statistical mechanics of complex networks, Rev. Modern Phys. 74 (1) (2002) 47–97, http://dx.doi.org/10.1088/1478-3967/1/3/006, arXiv:0106096v1.

[49] J. Kunegis, KONECT – the koblenz network collection, in: Proc. Int. Conf. on World Wide Web Companion, 2013, pp. 1343–1350, http://dl.acm.org/citation.cfm?id=2488173.

[50] T. Opsahl, P. Panzarasa, Clustering in weighted networks, Social Networks 31 (2009) 155–163, http://dx.doi.org/10.1016/j.socnet.2009.02.002.

[51] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, A. Arenas, Self-similar community structure in a network of human interactions, Phys. Rev. E 68 (2003) 065103, http://dx.doi.org/10.1103/PhysRevE.68.065103.

[52] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, W. Van den Broeck, What's in a crowd? analysis of face-to-face behavioral networks, J. Theoret. Biol. 271 (1) (2011) 166–180.

Robert Paluch received his B.Sc.Eng degree in Computer Physics in 2012 and M.Sc.Eng in Complex Systems Modeling in 2014 from Warsaw University of Technology, Poland. In 2012–2013 he worked for The European Organization for Nuclear Research (CERN) as a participant of the Technical Student Programme. Currently he is a Ph.D. student at the Warsaw University of Technology in the Group of Physics in Economy and Social Sciences. In years 2016–2019 he was visiting researcher at Rensselaer Polytechnic Institute.

Łukasz G. Gajewski is a Ph.D. student at the Warsaw University of Technology in the Group of Physics in Economy and Social Sciences. He has received his B.Sc. Eng degree in Computer Physics and M.Sc. Eng in Complex Systems Modeling. In years 2018–2019 he seconded at Stanford University and Rensselaer Polytechnic Institute.

Janusz A. Holyst is a Full Professor at Faculty of Physics, Warsaw University of Technology where he leads Group of Physics in Economy and Social Sciences. His current research includes simulations of evolving networks, models of collective opinion and emotion formation, and phase transitions. His list of publications includes around 150 papers in peer reviewed journals that have been cited almost 3000 times. He is currently one of Main Editors of Physica A.

Boleslaw K. Szymanski is the Claire and Roland Schmitt Distinguished Professor of Computer Science and the Founding Director of the Center for Network Science and Technology, Rensselaer Polytechnic Institute. He was the Director of the Social Cognitive Networks Center in the Network Science Collaborative Technology Alliance, the Principal Investigator in the International Technology Alliance, and in the Milky-Way@home project that models the distribution of dark matter in the Milky Way Galaxy. His projects focus on dynamic processes on networks, groups in social networks, sensor network protocols and algorithms, and large-scale distributed computing. Dr. Szymanski was a visiting professor at Universities of Pennsylvania, Stanford and Wrocłław University of Technology. He is a foreign member of Polish Academy of Sciences.