

VAIM: Visual Analytics for Influence Maximization*

Alessio Arleo¹[0000-0003-2008-3651], Walter Didimo²[0000-0002-4379-6059],
Giuseppe Liotta²[0000-0002-2886-9694], Silvia Miksch¹[0000-0003-4427-5703], and
Fabrizio Montecchiani²[0000-0002-0543-8912]

¹ TU Wien, Austria

`name.surname@tuwien.ac.at`

² Università degli Studi di Perugia, Italy

`name.surname@unipg.it`

Abstract. In social networks, individuals’ decisions are strongly influenced by recommendations from their friends and acquaintances. The *influence maximization* (IM) problem asks to select a *seed set* of users that maximizes the influence spread, i.e., the expected number of users influenced through a stochastic diffusion process triggered by the seeds. In this paper, we present VAIM, a visual analytics system that supports users in analyzing the information diffusion process determined by different IM algorithms. By using VAIM one can: (i) simulate the information spread for a given seed set on a large network, (ii) analyze and compare the effectiveness of different seed sets, and (iii) modify the seed sets to improve the corresponding influence spread.

Keywords: : Influence Maximization · Information Diffusion · Visual Analytics

1 Introduction

People in social networks influence each other in both direct and indirect ways, through a mechanism often known as the *word-of-mouth effect* (see, e.g., [11,12]). For this reason social networks are becoming the favorite venue where companies advertise their products/services and where politicians run their campaigns. The *influence maximization* (IM) problem asks to select a *seed set* of users that maximizes the influence spread, i.e., the expected number of users positively influenced by an information diffusion process triggered by the seeds and that spreads through the network according to some stochastic model. We refer the reader to the works by Guille et al. [9] and by Li et al. [15] for surveys about influence maximization and information diffusion in social networks.

* Research of WD, GL and FM partially supported by: (i) MIUR, grant 20174LF3T8 “AHeAD: efficient Algorithms for HARnessing networked Data”, (ii) Dip. di Ingegneria - Università degli Studi di Perugia, grant RICBA19FM: “Modelli, algoritmi e sistemi per la visualizzazione di grafi e reti”. Research of AA and SM partially supported by TU Wien “Smart CT” research cluster.

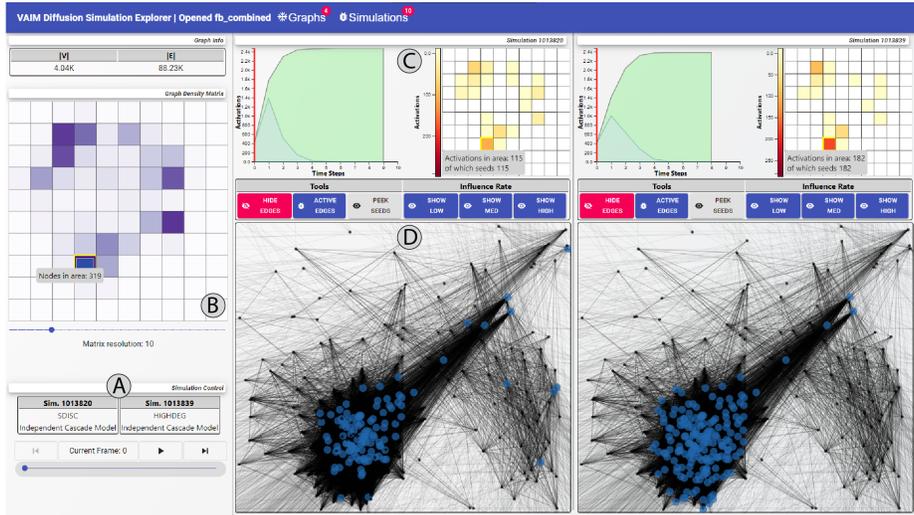


Fig. 1. VAIM’s visual interface, at $t = 0$ of the case study in Section 3. Its components are marked as follows: A) Simulation control, B) Density Matrix view, C) Diffusion Matrix view, D) Node-link view.

Analyzing and engineering an IM algorithm is a demanding task; as reported by Arora et al. [3], there is no single state-of-the-art technique for IM. Under the most common diffusion models, finding the optimal seed set in a network is known to be an NP-hard problem [11]. Besides the problem hardness, being the information diffusion process stochastic, even the evaluation of influence spread of any seed set is computationally complex [7], which makes the design of scalable and effective IM algorithms a great challenge that motivated a large and still increasing body of literature [15]. In this context, we want to exploit the power of information visualization to support expert users in analyzing, evaluating, and comparing IM algorithms. Our main contributions are as follows.

(i) We present VAIM, a system that provides facilities to simulate an information diffusion process over a given network and problem-oriented visual analytics (VA) tools to explore the related data (Section 2). VAIM has a modular architecture that currently includes some of the most popular IM algorithms and information diffusion models. An interface with multiple coordinated views makes it possible to visually compare and analyze the performance of a diffusion model over potentially very large networks and for different choices of the seed sets (i.e., for different IM algorithms). The user can interactively modify the seed set and iterate the process until a satisfying spread is achieved.

(ii) The effectiveness of VAIM is evaluated through a case study (Section 3). We show how tacking advantage of VAIM for (a) comparing different seed selection algorithms on the same network, and (b) improving the seed selection by either a manual or a system-assisted modification of the initial seed set.

Related work. There are several visualization systems designed to analyze information diffusion processes in social networks. TwitInfo [18,17] aggregates tweets in the spatial, temporal, and event dimensions supporting the exploration of event propagation processes. Whisper [4] exploits a flower-like visualization for real-time monitoring of the diffusion of a given topic, highlighting the spatio-temporal information of the process over the world. OpinionFlow [23] uses Sankey graphs and density maps to visually summarize opinion diffusion processes. FluxFlow [24] adopts a timeline visualization to analyze anomalous information diffusion spreading. D-Map [5] collects data from Sina Weibo and offers a map-based ego-centric visualization to reveal dynamic patterns of how people are involved and influenced in a diffusion process. SocialWave [20] uses abstract visualizations to explore and analyze spatio-temporal diffusion of information. More approaches are elaborated in Chen et al. [6]. All these approaches are designed to reveal different facets of information diffusion processes and they often rely on geographical and other user-related information. On the other hand, they neither support the user in analyzing the impact of the seeds (which in fact may be unknown) and of the network structure in terms of influence spread, nor offer simulation tools to experiment different diffusion models.

Long and Wong [16] introduce Visual-VM, a visualization system for viral marketing. Similar to VAIM, Visual-VM allows users to simulate stochastic diffusion processes and to visually analyze their output. However, Visual-VM offers a simple visual interface, which strongly relies on geographical information to lay out the network. The networks analyzed with VAIM may come from diverse scenarios and may not contain geographical information about users.

Finally, Vallet et al. [21,22] present a visualization framework to compare different diffusion models based on a common set of graph rewriting rules. Different from VAIM, the work of Vallet et al. does not focus on comparing different IM algorithms and it is mainly tailored to networks of small or medium size.

Background and notation. We model a social network as a directed graph $G = (V, E)$. A *diffusion model* M captures the stochastic diffusion process among the vertices of G . During the process, a vertex $v \in V$ can be either *active* or *inactive*. The *influence spread* of a seed set S , denoted by $\sigma_{G,M}(S)$, is the expected number of active vertices once the diffusion process (over the graph G and under the model M) terminates. More formally, the IM problem asks for a set $S^* \subseteq V$ of at most $0 < k \leq |V|$ seeds that maximizes the influence spread, i.e., $S^* = \text{argmax}\{\sigma_{G,M}(S) | S \subseteq V \wedge |S| \leq k\}$. One of the most commonly used diffusion models is the *Independent Cascade* (IC) [15]. Other models (such as the Linear Threshold model) make use of additional parameters but do not differ significantly in terms of the underlying iterative framework. In the IC model, a diffusion instance unfolds through an iterative process: In step 0, only the seed vertices are active; in step $j > 0$, each vertex u activated at step $j - 1$ will activate each of its inactive neighbors v with probability $0 \leq p(u, v) \leq 1$. The process halts when no more vertices can be activated. Unfortunately, the IM problem is NP-hard under the IC model, as well as under other models [11]. For a broader discussion refer to [9,15].

2 VAIM Design

The design of VAIM relies on the “Data-Users-Tasks” model proposed in [19].

Data. To estimate the influence spread of a seed set, we rely on a simulation-based approach. To obtain statistically relevant data, the simulation is repeated multiple times. Each single repetition is a time-dependent process taking as input a graph and a set of seeds. Hence, the data model of VAIM includes the input network, and set-typed temporal data represented by the active set of vertices and edges at every timestamp of the simulated diffusion process.

Users. VAIM targets a single class of expert users. Those users are knowledgeable in their own application domain and in the use of visual analytics tools. Also, they are interested not only in the resulting influence spread, but also on how the structure of the network influences the diffusion process.

Tasks. VAIM is designed to support the following user tasks:

T1 Simulate. It should be possible to simulate a diffusion process on a given network, with the seeds from an IM algorithm, under a given diffusion model.

T2 Evaluate. The user should be allowed to visually analyze both the quality of spread of a seed set and the impact of the network structure on the diffusion process, such as areas with a higher rate of active nodes, isolated areas, etc. The user can fast forward, rewind, and pause the process animation.

T3 Compare. It should be possible to visually compare the performance of different seed sets computed by different IM algorithms.

T4 Feedback. The user should be facilitated in modifying the seed set and iterate the simulate-evaluate-compare process.

2.1 Visualization design

The visualization design adopts an overview+detail approach. The interface is organized as a dashboard with multiple coordinated views (see also Fig. 1). The chosen colour schemes and palettes are colorblind friendly [10].

– **Simulation control (Fig. 1-A).** Here the user can set different parameters about the diffusion process, such as the stochastic model and the number of iterations (Task T1).

– **Density matrix view (Fig. 1-B).** The main purpose of this view is to provide an overview of the network structure in a scalable manner. This is achieved with a simplified matrix visualization, which is obtained by firstly computing a node-link layout of the whole (potentially very large) network with some fast algorithm, such as centralized or distributed force-directed techniques (e.g., [1,2,13]), and then by slicing the plane into cells. The color intensity of each cell reflects the number of nodes inside. The size of the matrix can be increased or decreased through a simple slider. Hovering with the mouse on a cell, the number of nodes in that cell is reported.

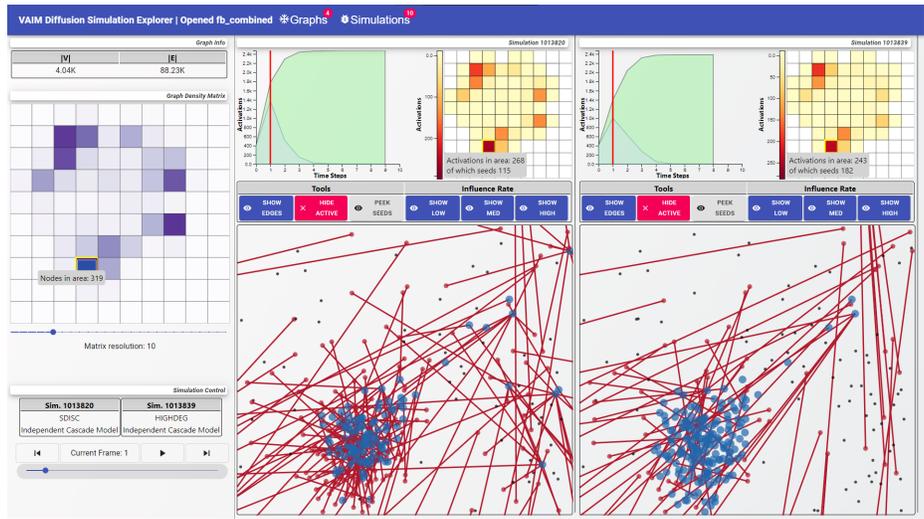
– **Diffusion matrix view** (Fig. 1-C). It allows users to visually compare multiple simulations over the same network. A legend below it shows the considered IM algorithms. Each simulation is conveyed using a distinct matrix visualization whose cells’ colors vary in a *YlOrRd* scale (yellow to orange to red) and reflect the number of active nodes in the corresponding area of the network. Notably, the density and diffusion matrices have the same set of cells, so to facilitate comparisons and associations among them. Similarly as for the density matrix view, the computation and the rendering of this view must be fast enough to allow the visualization of multiple simulations over large networks. At the left side of each diffusion matrix, the **process trend chart** is a plot with two curves showing, for each iteration, the number of new nodes activated in that iteration and the cumulative number of nodes activated up to that iteration. A red vertical segment indicates the currently selected iteration. VAIM can animate the diffusion process over time. Other facilities allow users to highlight those cells containing some seeds, or whose influence rate is low ($< 30\%$), medium ($[30\%, 60\%]$), or high ($> 60\%$). Clicking on a cell, its influence rate is shown and a list of nodes that can be either removed or promoted as seeds is suggested, based on node degrees and influence rate.

– **Node-link view** (Fig. 1-D). Below each diffusion matrix, there is a panel in which a detailed node-link diagram of a portion of the network can be visualized. This portion can be freely chosen by the user through a brushing selection of any group of $k \times h$ cells in the density matrix. The combination of the node-link view with the two matrix views described above is particularly useful for very large networks, for which detailed visualizations are feasible only for small portions. In the diagram, blue nodes represent seeds while dark red nodes and edges represent the active elements at the considered time instant (Fig. 2(a)). The user can hide all edges or leave only the active ones.

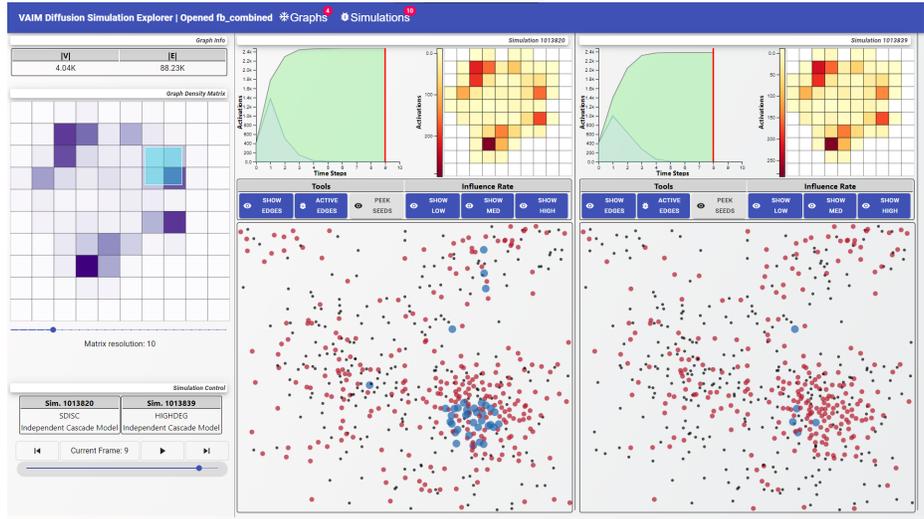
The three views together are designed to support Tasks T2, T3 and T4.

3 Evaluation and Discussion

We discuss an evaluation of VAIM based on the following case study (see the appendix for an additional case study). The input is the **fb-combined** social network, extracted from Facebook [14], having 4,039 nodes and 88,234 edges (<https://snap.stanford.edu/data/>). We simulated an IC diffusion process (T1), using two seed sets of 400 nodes each, computed by two popular IM algorithms: **HIGHDEG** [11,12] and **SDISC** [8], based on degree centrality and discount, respectively. We compared and evaluated (T2 and T3) the performance of the two diffusion processes. Fig. 1 shows a snapshot of the interface at the beginning of each diffusion process. The process trend charts reveal that **SDISC** leads to a higher number of active nodes in fewer iterations. By exploring the diffusion matrices we can observe a different distribution of the seeds selected by the two IM algorithms. For example, focusing on the densest cell of the network (which can be easily spotted in the density matrix), we see that **HIGHDEG** (the



(a)



(b)

Fig. 2. Snapshot of VAIM after (a) the first iteration of the diffusion process, and (b) at the end of the diffusion process.

right-side simulation) concentrates a higher number of seeds than SDISC (the left-side simulation) in that cell (182 seeds of HIGHDEG vs 115 seeds of SDISC), while putting relatively fewer seeds in sparser cells. Also, within the densest cell, SDISC distributes the seeds more uniformly than HIGHDEG. Fig. 2(a) shows the processes at the next iteration, and still focuses on the densest cell. Despite the smaller number of seeds, SDISC yields a higher number of newly active nodes (red nodes) in that cell (268 of SDISC vs 243 of HIGHDEG). Also, the greater number of red edges (those used by the diffusion process) exiting the cell, reveals a higher influence of the nodes of this cell towards nodes outside it. Fig. 2(b) shows the end of the processes. Using the influence rate function, we observe that the cells selected from the density matrix have a smaller number of active nodes with HIGHDEG than with SDISC. Looking at the node-link view for these cells (edges are hidden), this seems to be caused by the very small number of seeds that HIGHDEG placed in this portion of the network. The above discussion helps understanding how the seeding strategy adopted by SDISC leads to better performance, which corroborates the results of an experimental analysis performed on a collaboration graph presented in [8].

In order to improve the information spread of SDISC (T4), VAIM suggested 20 nodes (with smallest degree in the cell with highest influence rate) to be removed from the original seed set and 20 nodes (with highest degree in the cell with lowest influence rate) to be promoted as seeds. We modified the seed set accordingly and we simulated again the diffusion process. The new process lead to 2% more of active nodes.

4 Conclusion and Future Work

We discussed the use of visual analytics to support the analysis and fine tuning of IM strategies. We plan to extend the system with features such as edge bundling to mitigate edge clutter in the node-link view. We will also implement new diffusion models, together with ad-hoc views to explore the additional parameters of these models. Considering networks with node and edge attributes (e.g., geo-locations) is also an interesting direction. Finally, we want to further evaluate VAIM with more case studies and experiments, in particular to test its scalability (both in terms of simulation and visualization) to very large networks.

References

1. Arleo, A., Didimo, W., Liotta, G., Montecchiani, F.: Large graph visualizations using a distributed computing platform. *Inf. Sci.* **381**, 124–141 (2017). <https://doi.org/10.1016/j.ins.2016.11.012>, <https://doi.org/10.1016/j.ins.2016.11.012>
2. Arleo, A., Didimo, W., Liotta, G., Montecchiani, F.: A distributed multilevel force-directed algorithm. *IEEE Trans. Parallel Distrib. Syst.* **30**(4), 754–765 (2019). <https://doi.org/10.1109/TPDS.2018.2869805>, <https://doi.org/10.1109/TPDS.2018.2869805>
3. Arora, A., Galhotra, S., Ranu, S.: Debunking the myths of influence maximization: An in-depth benchmarking study. In: *SIGMOD Conference*. pp. 651–666. ACM (2017)
4. Cao, N., Lin, Y., Sun, X., Lazer, D., Liu, S., Qu, H.: Whisper: Tracing the spatiotemporal process of information diffusion in real time. *IEEE Trans. Vis. Comput. Graph.* **18**(12), 2649–2658 (2012)
5. Chen, S., Chen, S., Wang, Z., Liang, J., Yuan, X., Cao, N., Wu, Y.: D-Map: Visual analysis of ego-centric information diffusion patterns in social media. In: *VAST*. pp. 41–50. IEEE Computer Society (2016)
6. Chen, S., Lin, L., Yuan, X.: Social media visual analytics. *Comput. Graph. Forum* **36**(3), 563–587 (2017)
7. Chen, W., Wang, C., Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: *KDD*. pp. 1029–1038. ACM (2010)
8. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. p. 199208. *KDD 09*, Association for Computing Machinery, New York, NY, USA (2009). <https://doi.org/10.1145/1557019.1557047>, <https://doi.org/10.1145/1557019.1557047>
9. Guille, A., Hacid, H., Favre, C., Zighed, D.A.: Information diffusion in online social networks: a survey. *SIGMOD Record* **42**(2), 17–28 (2013)
10. Harrower, M., Brewer, C.A.: Colorbrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal* **40**(1), 27–37 (2003)
11. Kempe, D., Kleinberg, J.M., Tardos, É.: Maximizing the spread of influence through a social network. In: *KDD*. pp. 137–146. ACM (2003)
12. Kempe, D., Kleinberg, J.M., Tardos, É.: Maximizing the spread of influence through a social network. *Theory of Computing* **11**, 105–147 (2015). <https://doi.org/10.4086/toc.2015.v011a004>
13. Kobourov, S.G.: Force-directed drawing algorithms. In: Tamassia, R. (ed.) *Handbook on Graph Drawing and Visualization*, pp. 383–408. Chapman and Hall/CRC (2013)
14. Leskovec, J., Mcauley, J.J.: Learning to discover social circles in ego networks. In: *Advances in neural information processing systems*. pp. 539–547 (2012)
15. Li, Y., Fan, J., Wang, Y., Tan, K.: Influence maximization on social graphs: A survey. *IEEE Trans. Knowl. Data Eng.* **30**(10), 1852–1872 (2018)
16. Long, C., Wong, R.C.: Visual-VM: A social network visualization tool for viral marketing. In: *ICDM Workshops*. pp. 1223–1226. IEEE Computer Society (2014)
17. Marcus, A., Bernstein, M.S., Badar, O., Karger, D.R., Madden, S., Miller, R.C.: Processing and visualizing the data in tweets. *SIGMOD Record* **40**(4), 21–27 (2011)
18. Marcus, A., Bernstein, M.S., Badar, O., Karger, D.R., Madden, S., Miller, R.C.: TwitInfo: aggregating and visualizing microblogs for event exploration. In: *CHI*. pp. 227–236. ACM (2011)

19. Miksch, S., Aigner, W.: A matter of time: Applying a data–users–tasks design triangle to visual analytics of time-oriented data. *Computers & Graphics* **38**, 286–290 (2014)
20. Sun, G., Tang, T., Peng, T., Liang, R., Wu, Y.: Socialwave: Visual analysis of spatio-temporal diffusion of information on social media. *ACM TIST* **9**(2), 15:1–15:23 (2018)
21. Vallet, J., Kirchner, H., Pinaud, B., Melançon, G.: A visual analytics approach to compare propagation models in social networks. In: Rensink, A., Zambon, E. (eds.) *Proceedings Graphs as Models, GaM@ETAPS 2015*, London, UK, 11-12 April 2015. *EPTCS*, vol. 181, pp. 65–79 (2015). <https://doi.org/10.4204/EPTCS.181.5>, <https://doi.org/10.4204/EPTCS.181.5>
22. Vallet, J., Pinaud, B., Melançon, G.: Studying propagation dynamics in networks through rule-based modeling. In: Chen, M., Ebert, D.S., North, C. (eds.) *2014 IEEE Conference on Visual Analytics Science and Technology, VAST 2014*, Paris, France, October 25-31, 2014. pp. 281–282. IEEE Computer Society (2014). <https://doi.org/10.1109/VAST.2014.7042530>, <https://doi.org/10.1109/VAST.2014.7042530>
23. Wu, Y., Liu, S., Yan, K., Liu, M., Wu, F.: OpinionFlow: Visual analysis of opinion diffusion on social media. *IEEE Trans. Vis. Comput. Graph.* **20**(12), 1763–1772 (2014)
24. Zhao, J., Cao, N., Wen, Z., Song, Y., Lin, Y., Collins, C.: #FluxFlow: Visual analysis of anomalous information spreading on social media. *IEEE Trans. Vis. Comput. Graph.* **20**(12), 1773–1782 (2014)

Appendix

A Additional Case Study

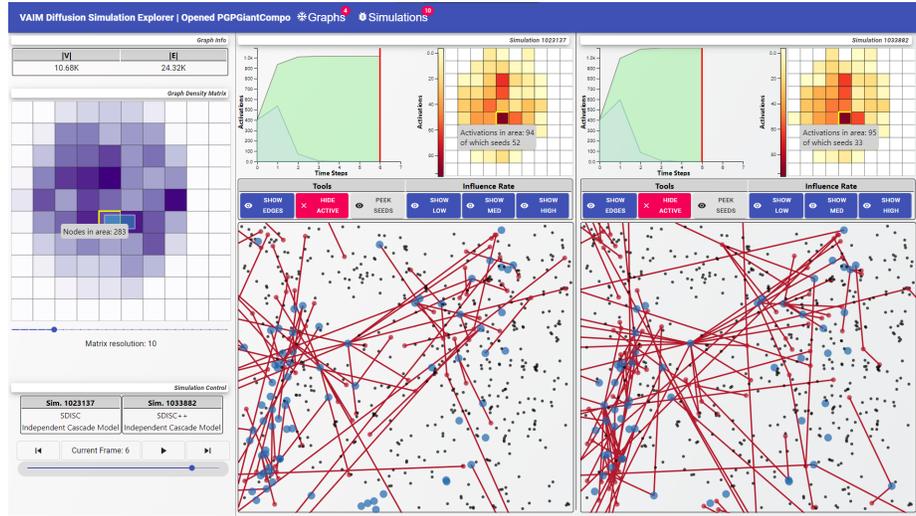


Fig. 3. Snapshot of VAIM at the end of the diffusion process for the second case study. The system suggested a change in the seed distribution (right side) over the one computed by SDISC (left side): in particular, removing seeds from the highlighted high-rate region did not harm the local spread performance, which increased by two units instead.

We performed a second case study on the email-exchange network `pgp-giant-compo` (<http://networkrepository.com/>), having $n = 10,680$ nodes and $m = 24,320$ edges. As in the first case study, we simulated an IC diffusion process with a seed set of 400 nodes computed by the SDISC IM algorithm.

By using the influence rate function, we identify the cell with the lowest number of activations. The system suggests a list of 20 nodes to add to the seed set, chosen among those with the highest degree located in the selected cell. Afterwards, we ask VAIM to provide a list of seeds to remove from the SDISC selection. These are picked among the nodes with lowest degree placed in the cell with the highest activation rate. We ran the simulation again with the modified seed set (that still holds the same number of elements) and we obtain an average increase in the spread performance of 1% (around 100 nodes).

In Figure 3 we report the comparison of the two diffusion processes at the end of the simulation. By looking at the two diffusion matrices, we could see how the updated seed set achieves higher activation rates on the upper left side of the network, where the new seeds were placed. Moreover, the seeds removed

from the most dense cell (in terms of number of activated nodes) did not harm the spread in that area; on the contrary, we observed a small increase in the number of activations.