Method of Graph Mining based on the Topological Anomaly Matrix and its Application for Discovering the Structural Peculiarities of Complex Networks

Artem Potebnia
Independent Investigator
Korosten, Ukraine
ORCID: 0000-0002-8162-5613

***Abstract*—The article introduces the mathematical concept of the topological anomaly matrix providing the foundation for the qualitative assessment of the topological organization underlying the large-scale complex networks. The basic idea of the proposed concept consists in translating the distributions of the individual vertex-level characteristics (such as the degree, closeness, and betweenness centrality) into the integrative properties of the overall graph. The article analyzes the lower bounds imposed on the items of the topological anomaly matrix and obtains the new fundamental results enriching the graph theory. With a view to improving the interpretability of these results, the article introduces and proves the theorem regarding the smoothness of the closeness centrality distribution over the graph's vertices. By performing the series of experiments, the article illustrates the application of the proposed matrix for evaluating the topology of the real-world power grid network and its post-attack damage.**

Keywords*—Topological anomaly matrix; complex network; graph topology; closeness centrality; betweenness centrality; power grid

I. Introduction

The distinctive feature characterizing the upcoming fourth wave of the industrial revolution lies in the rapid expansion, complication, and integration of the complex networks serving the needs of humanity and world economy [1, 2]. While enabling the development of the more efficient business processes leading to the increase in the produced outcome and quality of service, such tendency makes the entire society extremely vulnerable to the disruptions of the most critical infrastructural networks [3, 4]. Meanwhile, the functionality and reliability of any complex network heavily relies on its topology inspiring the emergent properties that could not be deduced from the separate network's entities and arise only in result of their interaction [5, 6]. For example, the United States of America has suffered from several catastrophic blackouts caused by the cascading failures in the power grid steaming largely from the low redundancy of its topological design [7 – 9]. These observations contribute to the particular reasonableness of assessing the topology of complex networks while making decisions regarding their reliability or need for the additional protection. Remark that this article focuses on considering the complex networks modeled by the undirected simple graphs $G=(V,E)$. In turn, the topology of any graph G could be regarded as the class of all possible graphs that are isomorphic to G. The evaluation of such topology is extremely challenging due to its underlying combinatorial nature and serves as a core problem of the emerging Big Data graph mining and analytics [10, 11]. In the prior works, the graph topology is assessed based on applying the quantitative metrics summarized in the review [12]. However, these metrics give a limited insight into the qualitative topological properties such as the concentration of bottlenecks inspiring the non-uniform load on the entities and links of the modeled network, which points to the presence of the research gap. Thereby, ***the objective of this article*** lies in constructing the mathematical object of the topological anomaly matrix providing the qualitative evaluation of the graph topology and its richness in bottlenecks, while satisfying the computational efficiency demands imposed to the instruments of the Big Data analytics.

II. Related Work: Vertex Importance Metrics

The inhomogeneous topology of graph gives rise to the differentiation in the relative importance of its nodes for ensuring the normal activity of the modeled complex network. However, the vertex importance is difficult for analyzing due to the possibility of its consideration from the radically different conceptual viewpoints. Thereby, in the existing works, the comprehensiveness of assessing the importance of the graph's nodes is ensured through applying a family of the formalized centrality metrics. In particular, the degree $d(v)$ of the vertex v reflects the extent of its local importance and serves as the simplest centrality metric. Nevertheless, the value of degree is incapable of capturing the position of the examined vertex within the entire graph. At the same time, the metrics of the closeness and betweenness centrality [13, 14] provide the formal way for evaluating the global importance of the graph's nodes and are defined in the following way:

Definition 1. The *closeness centrality* $c(v)$ of the node v belonging to the vertex set V of the connected graph G represents the inverted value of its average geodesic distance $d(v,\ k)$ to all nodes $k \in V \setminus \{v\}$, i.e.

$$c(v)=\frac{|V|-1}{\sum\limits_{k\in V\setminus\{v\}} d(v,\ k)}.$$

Definition 2. The *betweenness centrality* $b(v)$ reflects the likelihood that the examined vertex v appears on the shortest path between a pair of other nodes and is calculated as follows:

$$b(v) = \sum_{\substack{k,l \in V \setminus \{v\} \\ k \neq l}} \frac{\sigma_{kl}(v)}{\sigma_{kl}}.$$

Here σ_{kl} denotes the total number of the shortest paths between the vertices k and l that differ in at least one edge, while $\sigma_{kl}(v)$ stands for the number of such paths transiting the vertex v.

Intuitively, the closeness centrality could be interpreted as the velocity of the information broadcasting from the examined vertex to all other nodes of the graph. For example, by starting to spread from the nodes with the highest closeness centrality, the computer worms could potentially reduce the time required for infecting all vertices. For its part, the betweenness centrality could be viewed as the extent to which the examined vertex is involved as an intermediate in the communication flows between the other graph's nodes. Moreover, the vertices that ensure gluing together multiple implicit communities take the crucial responsibility for the exchange of information between them and, thereby, are typically characterized by the high betweenness centrality (especially in the case of the strong community structure) [15, 16].

III. Proposed Concept of the Topological Anomaly Matrix and its Fundamental Properties

The main contribution of this article lies in introducing the following mathematical object embodying the strategy of translating the local vertex-level characteristics into the property of the overall graph G:

Definition 3. The *topological anomaly matrix* $\mathbf{A}_{\mathbf{\Omega}}(G)$ of the graph G with respect to the *base vector* $\mathbf{\Omega} = [\omega_1 \ \ldots \ \omega_n]$ containing n vertex importance metrics $\omega_i : V \to \mathbf{R}$ is given in the form of the following $n \times n$ array:

$$\mathbf{A}_{\mathbf{\Omega}}(G) = \begin{bmatrix} a_{\omega_1}^{\omega_1}(G) & a_{\omega_2}^{\omega_1}(G) & \ldots & a_{\omega_n}^{\omega_1}(G) \\ a_{\omega_1}^{\omega_2}(G) & a_{\omega_2}^{\omega_2}(G) & \ldots & a_{\omega_n}^{\omega_2}(G) \\ \ldots & \ldots & \ldots & \ldots \\ a_{\omega_1}^{\omega_n}(G) & a_{\omega_2}^{\omega_n}(G) & \ldots & a_{\omega_n}^{\omega_n}(G) \end{bmatrix}.$$

Here the value of each item $a_{\omega_k}^{\omega_i}(G)$ lies within the range $[-1,1]$ and represents the bivariate correlation coefficient over all pairs of the set $DS(G, \omega_i, \omega_k) = \{(\omega_i(v), \omega_k(v)) | v \in V\}$. Note that $a_{\omega_k}^{\omega_i}(G)$ is taken to be undefined if either ω_i or ω_k is constant on the entire vertex set V (i.e. if there exists such $x \in \mathbf{R}$ that $\omega_i(V) \to \{x\}$ or $\omega_k(V) \to \{x\}$).

By definition, the matrix $\mathbf{A}_{\mathbf{\Omega}}(G)$ is symmetric, while its undefined components should be organized into the rows and columns crossing at the diagonal entries $a_{\omega_i}^{\omega_i}$ and, thereby, indicating the incapability of the corresponding metrics ω_i to distinguish the vertices of G. In turn, all defined components comprising the main diagonal of $\mathbf{A}_{\mathbf{\Omega}}(G)$ should be equal to one. For convenience, the matrices $\mathbf{A}_{\mathbf{\Omega}}(G)$ deprived of the undefined entries are referred to as *perfect* through this article.

The selection of metrics into the base vector $\mathbf{\Omega}$ is driven by essential need for ensuring the descriptiveness of the constructed matrix $\mathbf{A}_{\mathbf{\Omega}}(G)$ in assessing the topology of G at the optimal utilization of resources involved in the process of its calculation. In particular, the conceptual interpretability and computational efficiency of the metrics discussed in the previous section points to the reasonableness of introducing the *canonical base vector* defined as $\tilde{\mathbf{\Omega}} = [d \quad c \quad b]$. At the same time, the *canonical matrix* $\mathbf{A}_{\tilde{\mathbf{\Omega}}}(G)$ relying on such vector has the size of 3×3, while its full specification requires values of only three items $a_c^d(G)$, $a_b^d(G)$, and $a_b^c(G)$.

Remark that the matrices $\mathbf{A}_{\tilde{\mathbf{\Omega}}}(R)$ characterizing the purely random (and thereby unstructured) connected graphs R following the binomial distribution of vertex degrees tend to have the close-to-one values of all non-diagonal components. This tendency steams from the fact that, simply by chance, the higher-degree vertices demonstrate a larger probability of being located at the lower average distance to all other nodes and are likely to participate in the larger fraction of the shortest paths between them. In view of these considerations, every low (i.e. close-to-zero or negative) entry of the matrix $\mathbf{A}_{\tilde{\mathbf{\Omega}}}(G)$ clearly points to the significant non-randomness of the graph G and reveals the presence of the unexpected anomaly in its topology. In total, the matrix $\mathbf{A}_{\tilde{\mathbf{\Omega}}}(G)$ could encapsulate three major anomalies originating from the manner of fragmenting the graph G into the cohesive implicit communities.

In particular, the low value of $a_c^d(G)$ indicates that the larger number of the direct neighbors attached to an arbitrary vertex of G does not shrink its farness from the rest nodes of the graph to the statistically significant extent. The main topological property responsible for producing such anomaly consists in differentiating the entire communities of G into the central and peripheral ones (depending on the average distance to the other communities in terms of the inter-community edges). In this context, the high-degree vertices involved in the peripheral communities as well as the low-degree nodes occurring in the central ones serve as the key factors contributing to the reduce in the value of $a_c^d(G)$.

Conversely, the topological anomaly evidenced by the low value of $a_b^d(G)$ implies that the higher-degree vertices do not act as the significantly more preferred intermediates in the

shortest paths of the graph G. The topological pattern provoking such effect is characterized by the incidence of many critical inter-community edges to the low-degree nodes along with the presence of the high-degree vertices adjacent exclusively to the members of their own communities. Finally, at the low value of $a_b^c(G)$, the ability of an arbitrary vertex to be involved into the shortest paths in the graph G (and control the corresponding communication flows) is not strongly dependent on its average distance to the other vertices. From the topological viewpoint, the anomalous decrease in $a_b^c(G)$ is driven by the nodes that, while being located in the central communities, are neither directly incident to the inter-community edges nor lie on the shortest path between any pair of vertices equipped with such edges.

In order to provide a fruitful insight into the entries of $\mathbf{A}_{\tilde{\Omega}}(G)$, let us introduce and prove the following fundamental relationship between the closeness centrality values of the adjacent graph's nodes:

Theorem 1. The closeness centrality $c(v)$ of any vertex v in the connected graph $G=(V,E)$ is bounded below by

$$c(v)\geq\frac{|V|-1}{\frac{|V|-1}{c_m(v)}+|V|-2},$$

where $c_m(v)=\max\{c(u)|(v,u)\in E\}$ stands for the highest closeness centrality among all direct neighbors of v.

▲ Let us assume that v is adjacent to the node u having the closeness centrality of $c(u)$. This, for its part, implies that every vertex $h\in V\setminus\{v,u\}$ could be reached from v based on the walk (i.e. sequence of edges with allowed repetitions) composed of the edge (v,u) and shortest path from u to h. Accordingly, the geodesic distance between v and h is bounded above by the condition $d(v,h)\leq d(u,h)+1$, while $d(v,u)=1$. In view of this observation, the entire closeness centrality of v is constrained in the next manner:

$$c(v)\geq\frac{|V|-1}{|V|-2+\theta(v,u)};\quad \theta(v,u)=1+\sum_{h\in V\setminus\{v,u\}}d(u,h).$$

In turn, $\theta(v,u)$ could be expressed based on the closeness centrality of u as $\theta(v,u)=(|V|-1)/c(u)$, which completes deriving the desired relationship. At the same time, the increase in $c(u)$ over the whole allowed range $(0,1]$ leads to the monotonic growth of the imposed bound at any fixed $|V|\geq 3$. This remark clearly points to the largest restrictiveness of the bound produced by the neighbor with the highest closeness centrality. ▼

The most significant implication of the above theorem lies in the smooth nature of distributing the closeness centrality values over the graph's vertices. On the contrary, the values of the betweenness centrality could be distributed in much more rugged manner implying the extreme differences between the adjacent nodes. For example, each leaf vertex l, by definition, is associated with zero betweenness centrality $b(l)=0$ regardless the properties of its single neighbor. Conversely, the closeness centrality of l takes the lowest possible value satisfying the bound given in Theorem 1. Remark that such bound demonstrates the close-to-linear behavior at the low values of $c_m(v)$ (since its derivative with respect to $c_m(v)$ approaches one as $c_m(v)\to 0$). This observation clearly shows that the leaf nodes of the sparse large-scale graph G typically tend to have almost the same closeness centrality as their neighbors. In view of such relationship, the leaf vertices appearing in the central communities are characterized by the relatively high closeness centrality compared to the other graph's nodes and, thereby, serve as the most evident contributors to the reduce in the value of $a_b^c(G)$.

IV. Analysis of the Lower Bounds Imposed on the Entries of the Canonical Topological Anomaly Matrix

Meanwhile, the anomalous effects indicated by the matrix $\mathbf{A}_{\tilde{\Omega}}(G)$ are not inspired solely by the intentional self-organizing process of the complex network modeled by the graph G. Additionally, the values of $a_c^d(G)$, $a_b^d(G)$, and $a_b^c(G)$ are affected by the structural constraint taking the form of the vertex degree multiset $D(G)=\{d(v)|v\in V\}$ containing the degrees of all nodes in G. Each multiset $D(G)$, for its part, characterizes the family $\Gamma_{D(G)}$ composed of all non-isomorphic graphs $G'\in\Gamma_{D(G)}$ such that $D(G')=D(G)$. In this sense, the specification of $D(G)$ restricts the possible topologies of G only to ones contained in $\Gamma_{D(G)}$ and imposes the structural bounds on the components of $\mathbf{A}_{\tilde{\Omega}}(G)$.

Furthermore, such quantitative characteristics of G as the order $|V|$ and density $\varphi(G)=2|E|/(|V|(|V|-1))$ are derived from $D(G)$ and by themselves provide the lower bounds on the items of $\mathbf{A}_{\tilde{\Omega}}(G)$. For convenience, let us denote the minimum values of $a_c^d(G)$, $a_b^d(G)$, and $a_b^c(G)$ over all graphs G containing $|V|$ nodes and having the density of $\varphi(G)$ respectively by $m_c^d(|V|,\varphi(G))$, $m_b^d(|V|,\varphi(G))$, and $m_b^c(|V|,\varphi(G))$. Notice that all these lower bounds are defined over the domain restricted by $\varphi_{tree}(|V|)\leq\varphi(G)<1$, where

$\varphi_{tree}(|V|) = 2(|V|-1)/(|V|(|V|-1))$. Such restriction steams from the impossibility of constructing any connected graph sparser than a tree along with the presence of only undefined items in the matrix $\mathbf{A}_{\tilde{\Omega}}(K)$ of each complete graph K having all possible edges.

With a view to simplifying the discussion of the results given in Fig. 1, let us use the notations $m_c^d(\varphi(G))|_k$, $m_b^d(\varphi(G))|_k$, and $m_b^c(\varphi(G))|_k$ for the dependences of $m_c^d(|V|,\varphi(G))$, $m_b^d(|V|,\varphi(G))$, and $m_c^b(|V|,\varphi(G))$ on $\varphi(G)$ at the value of $|V|$ fixed to k (representing slices of the illustrated surfaces). As evident from Fig. 1a, the dependence $m_c^d(\varphi(G))|_k$ for any considered k exhibits a single minimum located close the lowest allowed density $\varphi_{tree}(k)$. Moreover, such minimum becomes deeper with the increase in k, which is directly attributed to the growing number of possible topologies. Another notable feature of the analyzed surface consists in the presence of the wide plateau-like region where $m_c^d(|V|,\varphi(G))$ takes the close-to-one values. While being located at the high density $\varphi(G)$, this region complies with the limited suitability of the dense graphs to the elaboration the high-modular topology underlying the emergence of the structural anomalies. Conversely, the tree graphs could be strongly segregated into the sparse implicit communities, which acts as an explanation for the relatively low values of $m_c^d(|V|,\varphi_{tree}(|V|))$. However, the requirement regarding the sparsity of communities also hinders the formation of the structural anomalies. Accordingly, the minima of all considered dependences $m_c^d(\varphi(G))|_k$ are slightly deviated from $\varphi_{tree}(k)$. For example, Fig. 1a depicts the graph G_7 responsible for producing the minimum of $m_c^d(\varphi(G))|_7$. Remark that this graph implies the inclusion of the three-degree vertices into the peripheral communities (represented by cycles) and placement of the two-degree node as the connector between these communities. As a result, such connector is associated with the largest closeness centrality compared to all other vertices. The graphs on six or fewer nodes, in turn, could not contain the lower-degree vertex characterized by the larger closeness centrality than the higher-degree one due to the influence of the structural restrictions.

At the same time, the shape of the surfaces constructed in Figs. 1b and 1c requires the more careful investigation. The distinctive feature expressed by the experimentally registered dependences $m_b^d(\varphi(G))|_k$ and $m_b^c(\varphi(G))|_k$ consists in the presence of multiple local minima whose number grows with the increase in k (one at $k=4$ and $k=5$, two at $k=6$, and three at $k=7$). Remark that for every considered k, the local minima of both $m_b^d(\varphi(G))|_k$ and $m_b^c(\varphi(G))|_k$ are exhibited at the identical graph topologies and same values of $\varphi(G)$. Furthermore, the presented results allow noticing that the bounds $m_b^d(|V|,\varphi(G))$ and $m_c^b(|V|,\varphi(G))$ are lower than $m_c^d(|V|,\varphi(G))$ at the intermediate density $\varphi(G)$. These effects are fully attributable to the fact that the betweenness centrality is capable of producing the rugged distributions over the graph's nodes, while the closeness centrality is unavoidably subjected to the smoothing requirement proved in Theorem 1.

The inspection of the callouts in Fig. 1 shows that the topologies underlying the local minima of $m_b^d(\varphi(G))|_k$ and $m_b^c(\varphi(G))|_k$ are characterized by the presence of the densely interconnected group of the highest-degree vertices along with the inclusion of the low-degree nodes into the chain-like substructures. Moreover, the collected results allow discovering that the formation of such topologies is driven by the hidden fundamental rules. In particular, each graph labeled in Fig. 1 as G_1^k for $k \in \{5,6,7\}$ could be obtained based on constructing the diamond graph (i.e. complete graph on four vertices with one removed edge) with the subsequent linking of its two-degree nodes by the path containing $k-3$ edges. Each graph labeled as G_2^k for $k \in \{6,7\}$, in turn, contains such basic substructures as the three-length cycle C_3 and star S_{k-3} represented by a tree with $k-4$ leaf vertices. Its formation involves placing all possible edges between the nodes of C_3 and $S_{k-3} \setminus \{r\}$, where r denotes the central node of the star S_{k-3}. These trends suggest that the additional local minima arising in the dependences $m_b^d(\varphi(G))|_k$ and $m_b^c(\varphi(G))|_k$ with the increase in k are caused by the graph topologies following the new fundamental rules.

V. Application of the Proposed Matrix for Assessing the Topology of the Power Grid Network and its Post-Attack Damage

The role of this section lies in demonstrating the descriptive potential of the introduced mathematical structure in evaluating the qualitative topological properties of the real-world complex networks. As a sample dataset for investigation, this work uses the benchmark model of the power grid infrastructure of the United States of America available at the open-access network collection [17] and given by the undirected graph $P = (V_P, E_P)$. Notice that this graph is connected and contains 4 941 vertices reflecting the facilities responsible for producing and distributing electricity along with 6 594 edges modeling the high-voltage transmission lines. The canonical matrix $\mathbf{A}_{\tilde{\Omega}}(P)$ calculated for the described graph P is given by $a_c^d(P) = 0.2306$, $a_b^d(P) = 0.2766$, and $a_b^c(P) = 0.3536$. These values indicate the involvement of all considered structural anomalies in the topological organization of P, which serves as the natural result for the spatially distributed technological man-made system needing the constant supervision for preserving the desired functionality.

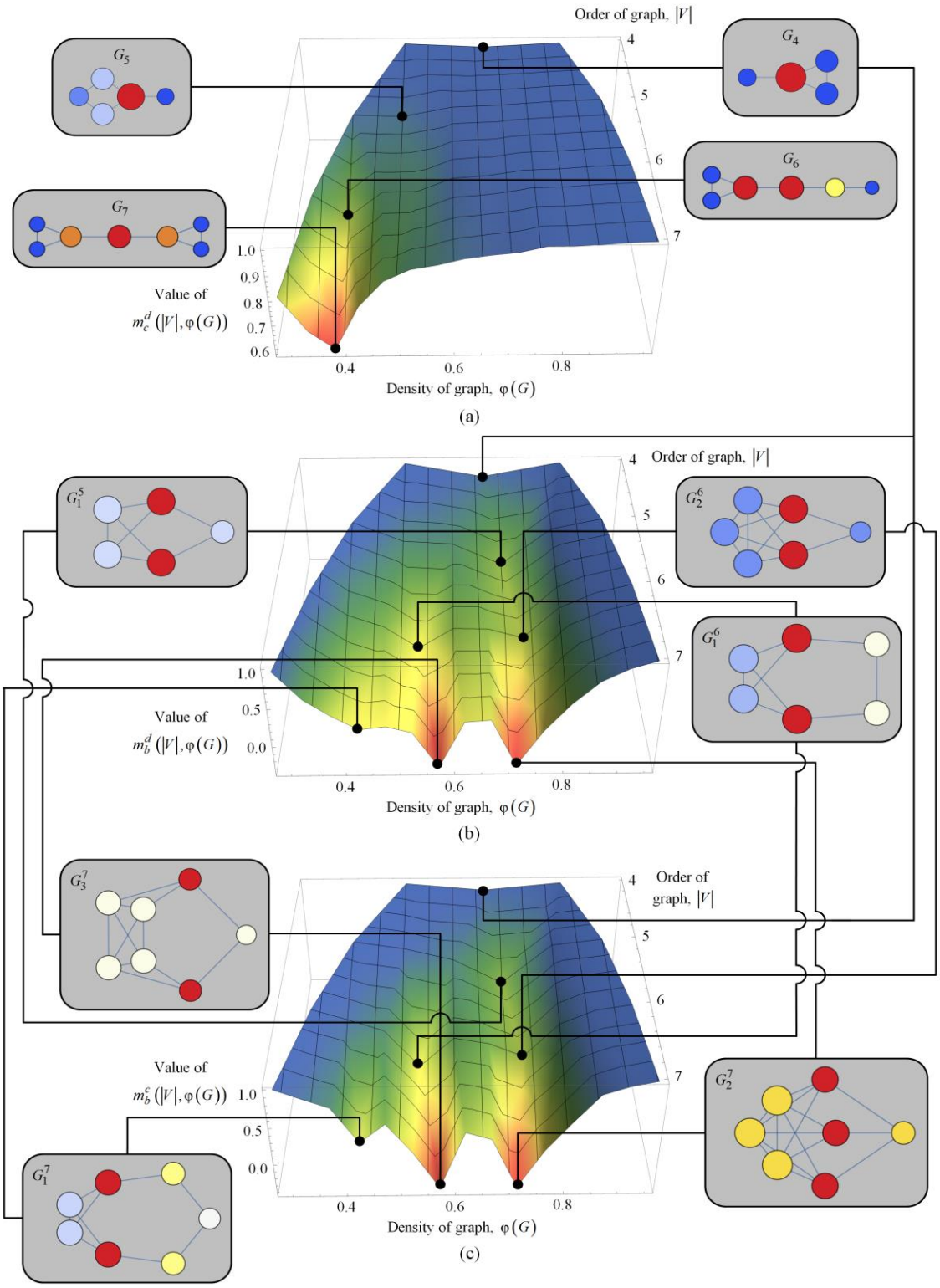

Fig. 1. Three-dimensional plots showing the lower bounds $m_c^d(|V|,\varphi(G))$, $m_b^d(|V|,\varphi(G))$, and $m_b^c(|V|,\varphi(G))$ on entries of the matrix $\mathbf{A}_{\tilde{\Omega}}(G)$ as functions of the graph order $|V|$ and density $\varphi(G)$. All given continuous surfaces were constructed by processing the scattered points representing the experimentally calculated data with the bilinear interpolation scheme. In addition, the surfaces are equipped with the callouts depicting the graph topologies underlying the minima of the considered bounds each fixed $|V|$. The size size and color of vertices in every callout graph encode respectively their closeness and betweenness centrality (the largest size and red color correspond to the highest values of metrics).

The more in-depth analysis shows that the sets of pairs $DS(P,d,c)$ and $DS(P,d,b)$ underlying the calculation of the correlation coefficients $a_c^d(P)$ and $a_b^d(P)$ are organized in line with inverted cone-shaped pattern of heteroscedasticity implying the decrease in the variability of the closeness and betweenness centrality values of nodes with increasing their degree. Remark that the emergence of such phenomenon steams heavily from the vertex degree distribution of P pointing to the presence of fewer nodes accommodating more neighbors. In turn, the set $DS(P,c,b)$ constituting the basis for computing the component $a_b^c(P)$ exhibits the direct cone-shaped form of heteroscedasticity characterized by the tendency of vertices with the higher closeness centrality to demonstrate the larger variability of the betweenness centrality values.

With a view to illustrate the usefulness of applying the proposed matrix as a measure of the topological damage, let us consider the attack on P implying the removal of all its nodes having the degree of at least t, i.e. comprising the subset $\Psi(t)=\{v \mid d(v)\geq t\}$. The post-attack graph on the remaining nodes of $V_P^t = V_P \setminus \Psi(t)$ is represented by $P_t=(V_P^t,E_P^t)$, where $E_P^t = E_P \setminus \{(v,v') \mid (v\in\Psi(t))\vee(v'\in\Psi(t))\}$. In turn, let us use the notation $W_t=(V_W^t,E_W^t)$ for the largest connected component of P_t. At the high fraction $f(t)=|V_W^t|/|V_P^t|$, W_t is additionally referred to as the giant component of P_t, while its topology accumulates the majority of damage that is not related to the connectivity issues [3].

Fig. 2 illustrates the application of the matrix $\mathbf{A}_{\tilde{\Omega}}(W_t)$ for assessing such damage by presenting the experimentally calculated values of its components $a_c^d(W_t)$, $a_b^d(W_t)$, and $a_b^c(W_t)$ as functions of the degree threshold t. Remark that with the decrease in t, all obtained dependences demonstrate the tendency to fall after the plateau-like region and reach their global minima at the same critical threshold $t_c=6$. Meanwhile, all post-attack graphs P_t for $t<t_c$ are deprived of the giant connected component (as evidenced by the dependence of $f(t)$ on t), while their subgraphs W_t are trees. In light of these observations, the sharp growth of $a_c^d(W_t)$, $a_b^d(W_t)$, and $a_b^c(W_t)$ at the subsequent reduce in t is driven by the structural constraints studied in the previous section. For its part, the graph W_{t_c} corresponding to the global minima of the traced dependences is characterized by the most significant anomalies reflecting the accumulation of the largest topological damage. Conceptually, with the decrease in t, such damage stimulates the collapse of W_{t_c} into the numerous small connected components. Furthermore, the dependence of $a_c^d(W_t)$ on t demonstrates the deepest global minimum $a_c^d(W_{t_c})=0.0784$. This result allows noting that the topological damage of W_{t_c} is expressed primarily by the more significant differentiation of its communities into the central and peripheral ones, which follows from the degradation of the inter-community relationships.

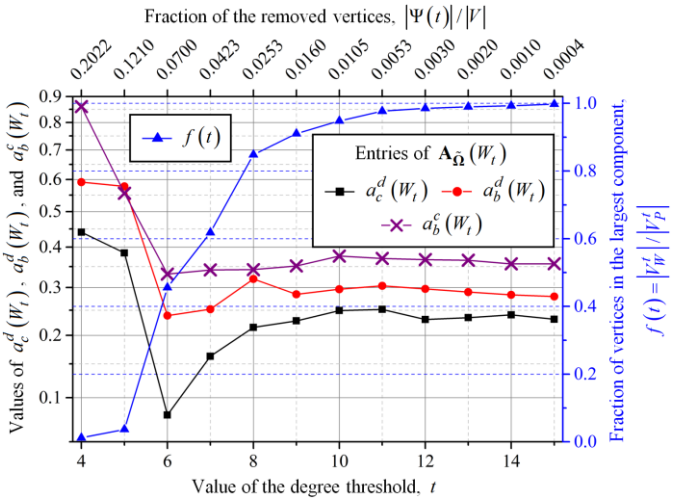

Fig. 2. Entries of the matrix $\mathbf{A}_{\tilde{\Omega}}(W_t)$ associated with the largest connected component W_t of the post-attack graph P_t obtained by deleting all vertices with the degree of at least t from the graph P modeling the power grid network. To trace the significance of the obtained results for the overall graph P_t, the plot additionally gives the fraction of its vertices included in W_t.

VI. Conclusions

The findings presented in the preceding sections clearly substantiate the crucial role of the topological anomaly matrix $\mathbf{A}_{\Omega}(G)$ in discovering the unexpected topological patterns of the real-world complex networks and producing the new fundamental results advancing the frontiers of the graph theory. The canonical form of the proposed matrix $\mathbf{A}_{\tilde{\Omega}}(G)$ is recommended for the widespread usage, while the need for performing the more in-depth analysis could be addressed by applying the matrices of larger size relying on the extended base vectors containing additional metrics. Conceptually, at the low values of $a_c^d(G)$ and $a_b^d(G)$, the network modeled by the graph G is characterized by the tendency of the entities accommodating only a few neighbors to act as hubs managing the significant portions of traffic. In turn, the links attached to such entities are subjected to the enhanced risks of overloading and, thereby, play a role of the primary structural bottlenecks. Meanwhile, the low value of $a_b^c(G)$ is caused by the entities that, while being located close to all other nodes, do not use their beneficial geodesic position to support the traffic transmission in the network and, in this sense, contribute to the formation of bottlenecks. In sum, the opportunity of ensuring the balance between the descriptive potential and computational complexity of the matrix $\mathbf{A}_{\Omega}(G)$ (through the

selection of metrics into the base vector **Ω**) allows its consideration as the promising tool in the Big Data graph mining and analytics. Due to its usefulness in describing the topological damage (as illustrated in the previous section), the topological anomaly matrix could be potentially applied as one of the robustness metrics in assessing the attack tolerance of complex networks. Similarly, the proposed matrix could assist in detecting the differences in the topological organization between the whole network and its important subnetworks (such as the rich-clubs). Another possible application lies in tracing the evolutionary topological transformation of complex networks (by comparing the matrices calculated for the giant connected components of the graph models constructed for the series of the time-indexed network snapshots).

REFERENCES

[1] M. Khan, X. Wu, X. Xu, and W. Dou, "Big Data challenges and opportunities in the hype of industry 4.0," in *Communications (ICC)*, 2017 IEEE International Conference on, pp. 1 – 6, 2017.

[2] Y. Lu, "Industry 4.0: A survey on technologies, applications and open research issues," *Journal of Industrial Information Integration*, vol. 6, pp. 1 – 10, 2017.

[3] A. Potebnia, "Innovative metrics for assessing the catastrophic collapse of the complex networks under the greedy attacks on their most important vertices and edges," in *Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET)*, 14th International Conference on, pp. 564 – 569, 2018.

[4] A. Potebnia, "Innovative concept of the strict line hypergraph as the basis for specifying the duality relation between the vertex separators and cuts," in *Advances in Intelligent Systems and Computing II. CSIT 2017. Advances in Intelligent Systems and Computing*, vol. 689, Springer International Publishing, pp. 386 – 403, 2018.

[5] Y. Huang, G. Wang, and Y. Tang, "Bottleneck attack strategies on complex communication networks," in *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence. ICIC 2010. Lecture Notes in Computer Science*, vol. 6216. Springer-Verlag Berlin Heidelberg, pp. 418 – 425, 2010.

[6] A. Potebnia, "New method for estimating the tree-likeness of graphs and its application for tracing the robustness of complex networks," in *Computing, Communication and Networking Technologies (ICCCNT)*, 9th International Conference on, 2018.

[7] Y.-K. Wu, S.M. Chang, and Y.-L. Hu, "Literature review of power system blackouts," *Energy Procedia*, vol. 141, pp. 428 – 431, 2017.

[8] S. Soltan, D. Mazauric, and G. Zussman, "Analysis of failures in power grids," *IEEE Transactions on Control of Network Systems*, vol. 4(2), pp. 288 – 300, 2017.

[9] L. Liu, Y. Yin, Z. Zhang, and Y.K. Malaiya, "Redundant design in interdependent networks," *PLoS ONE*, vol. 11(10):e0164777, 2016.

[10] J.A. Miller, L. Ramaswamy, K.J. Kochut, and A. Fard, "Research directions for Big Data graph analytics," in *Big Data (BigData Congress)*, 2015 IEEE International Congress on, pp. 785 – 794, 2015.

[11] M.U. Nisar, A. Fard, and J.A. Miller, "Techniques for graph analytics on Big Data", in *Big Data (BigData Congress)*, 2013 IEEE International Congress on, pp. 255 – 262, 2013.

[12] B. Kantarci and V. Labatut, "Classification of complex networks based on topological properties," in *Cloud and Green Computing (CGC)*, 2013 Third International Conference on, pp. 297 – 304, 2013.

[13] K. Das, S. Samanta, and M. Pal, "Study on centrality measures in social networks: a survey," *Social Network Analysis and Mining*, vol. 8:13, 2018.

[14] N. Matas, S. Martincic-Ipsic, and A. Mestrovic, "Comparing network centrality measures as tools for identifying key concepts in complex networks: A case of Wikipedia," *Journal of Digital Information Management*, vol. 15(4), pp. 203 – 213, 2017.

[15] S. Fortunato and C. Castellano, "Community structure in graphs," in *Computational Complexity*. Springer New York, pp. 490 – 512, 2012.

[16] G. Lin, Z. Di, and Y. Fan, "Cascading failures in complex networks with community structure," *International Journal of Modern Physics C*, vol. 25(5): 1440005, 2014.

[17] US power grid network dataset. KONECT: The Koblenz Network Collection. Available online: http://konect.uni-koblenz.de/networks /opsahl-powergrid. Accessed: April 2018.