A Graph-Based Method for Detecting Rare Events: Identifying Pathologic Cells

Enikö Székely, Arnaud Sallaberry, Faraz Zaidi, Pascal Poncelet

A Graph-based Method to Detect Rare Events: An Application to Identify Pathologic Cells

Enikö Székely
CIMS, New York University, USA

Arnaud Sallaberry
LIRMM, Université Paul Valéry Montpellier 3, France

Faraz Zaidi
University of Lausanne, Switzerland, and Karachi Institute of Economics and Technology, Pakistan

Pascal Poncelet
LIRMM, Université Montpellier 2, France

Abstract Detection of outliers and anomalous behavior is a well-known problem in the data mining and statistics fields. Although the problem of identifying single outliers has been extensively studied in the literature, little or some effort has been devoted to the detection of small groups of outliers that are similar to each other but markedly different from the entire population. Many real world scenarios have small groups of outliers, e.g. a group of students that excel in a classroom or a group of spammers in an online social network. In this paper, we propose a novel method to solve this challenging problem that lies at the frontiers of outlier detection and clustering of similar groups. The method transforms a multidimensional dataset into a graph, applies a network metric to detect clusters and renders a representation for visual assessment to find rare events. We test the proposed method to detect pathologic cells (e.g. Cancer, HIV, CVA, etc.) in the biomedical science domain. The results are very promising and confirm the available ground truth provided by the domain experts.

Keywords Outlier Detection, Rare Events, Group of Outliers, Visualization, Clustering, Pathologic Cells.

1 Introduction

With recent technological developments, the acquisition and storage of large datasets from various domains are now common. Outliers in these datasets can arise due to changes in system behavior, instrument or human error, intentional fraudulent behavior, or through natural deviations in population due to epidemics and virus infections. Detection of these outliers now has many applications regarding data cleansing, stopping fraudulent intentions, controlling disease outbreaks and detecting infected individuals, etc.

A recent outlier detection research area concerns the identification of a group of outliers, called rare events. Members of this group are similar to each other, but deviate from the general behavior of the dataset, thus arousing suspicions that they were generated by a different mechanism. Furthermore, these groups are usually very small compared to clusters or groupings within the entire dataset, so they are thus classified as outliers. Detection of such groups is more challenging when the aim is to only detect groups of such outliers that are similar to each other while ignoring outliers that are different amongst themselves. Examples of such rare events include identification of students in a class that excel academically or a group of spammers or autobots that increase the popularity of an individual or an event in an online social network.

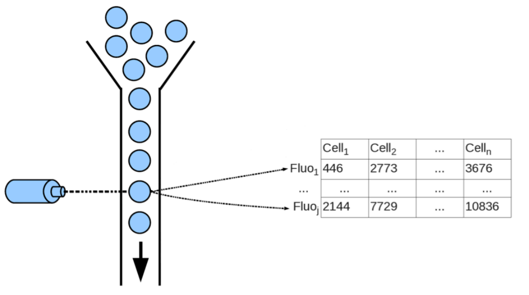
One critical application of detecting rare events arises in biomedical science and clinical research, e.g. the problem of extracting rare events from flow cytometry standard (FCS) data. Such files consists of multiparametric descriptions of thousands to millions of individual and rare cells, called biological markers of interest, and are used in monitoring vascular diseases, oncology and infectious diseases.

Detection of rare events with a high recall, i.e. with no false negatives, is critical in this domain as the cost of missing pathologic cells is significantly much higher than the cost of misclassifying a healthy group of cells.

In this article, we address this challenging problem of detecting rare events in FCS files. The proposed approach is based on initial candidate selection using kNN, filtering irrelevant candidates, applying a metric for detecting densely connected data items and rendering a representation for interactive visual detection of the cluster of interest. Results demonstrate the accuracy and effectiveness of the proposed algorithm when compared with available ground truth. Experiments were conducted to detect pathologic cells and we intend to extend this study to other types of datasets as part of future work.

Flow Cytometry

Flow cytometry is a technology that allows measurement of blood cell characteristics at very high rates (up to thousands of cells per second). Figure 1 gives an illustration of the flow cytometry process.



	Cell_1	Cell_2	...	Cell_n
Fluo_1	446	2773	...	3676
...	...	...	...	...
Fluo_j	2144	7729	...	10836

Figure 1: Principle of a flow cytometer. Cells are suspended in a stream and passed by a laser beam. The light hitting the cells is re-emitted according to the cell characteristics. Values, i.e. fluorescent signals from individual cells, are stored in corresponding fields for further analysis.

Each cell passes through one or more light beams which measures fluorescent signals from individual cells, i.e. different possible responses depending on the fluorophores that have been added to the blood sample. This fluorescence process generates considerable information about cells, and allows their separation (an antibody is linked to a fluorescent dye and bound to a protein that is discriminative between cells). Finally, fluorescence levels in response to cell markers are stored in flow cytometry standard (FCS) data files [1].

Current flow cytometers can count up to tens of millions of cells in normal cell populations found in any healthy patient, such as lymphocytes or monocytes. In patients presenting with a blood pathology, the blood samples also contain micro-clusters of cells with abnormal signatures, i.e. abnormal combinations of cell marker fluorescence levels.

The operator usually performs a visual detection by sequentially inspecting two-dimensional spaces, i.e. combinations of two markers (see Figure 2). This approach leads to very high inter-variability (17–44%) [2] among research laboratories regarding what defines an abnormal cell population and is sensitive to complex multivariate relationships.

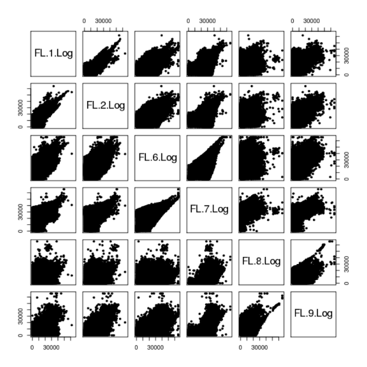


Figure 2: Original data events on a flow cytometry blood sample.

1. H. M. Shapiro. Practical flow cytometry: Wiley-Blackwell, 4th edition, 2003.
2. A. Bashashati and R. R. Brinkman, A survey of flow cytometry data analysis methods, Advances in Bioinformatics, pp. 1–19, 2009.

Defining an Outlier and a Group of Outliers

According to [1], “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”. [2] defines an outlier as “An observation or subset of observations which appear to be inconsistent with the remainder of that dataset”.

Other terminologies commonly used for outliers are novelty detection, anomaly detection, one-class classification, noise detection, deviation detection or exception mining [3]. Outlier detection has been found to be useful in numerous applications such as intrusion detection in computer networks, fraudulent usage of credit cards, topic detection in news documents and web pages, discovery of temporal changes in evolving online social networks, and identification of inconsistent digital records.

Analogous to an outlier, a group of outliers can be defined as a sub-population of individuals deviating from the general behavior of the entire population but similar to each other. The cardinality of this set of rare events is usually very small as compared to the general grouping of the dataset, which classifies them as outliers. We use the term ‘rare events’ in this text in reference to this group of outliers, whereas synonyms such as cluster of outliers [4], clustered anomaly [5], anomaly collection [6] and micro-clusters [7] are also used in the literature.

1. D. M. Hawkins. Identification of outliers, volume 11. Chapman and Hall London, 1980.
2. V. Barnett and T. Lewis. Outliers in statistical data. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, Chichester: Wiley, 1984, 2nd ed., 1, 1984.
3. V. Hodge and J. Austin. A survey of outlier detection methodologies. Artificial Intelligence Review, 22(2): 85-126, 2004.
4. D. M. Rocke and D. L. Woodruff. Identification of outliers in multivariate data. Journal of the American Statistical Association, 91(435): 1047-1061, 1996.
5. F. T. Liu, K. M. Ting, and Z.-H. Zhou. On detecting clustered anomalies using sciforest. In Machine Learning and Knowledge Discovery in Databases, pages 274-290. Springer, 2010.
6. H. Dai, F. Zhu, E.-P. Lim, and H. H. Pang. Mining coherent anomaly collections on web data. In Proceedings of CIKM, 1557-1561, 2012.
7. D.-H. Bae, S. Jeong, S.-W. Kim, and M. Lee. Outlier detection using centrality and center-proximity. In Proceedings of CIKM, 2251-2254, 2012.

Outlier Detection Approaches

There are different approaches to detect outliers in the literature [1]. We can determine outliers without any prior knowledge. This approach processes data and identifies the most distant points as outliers. If the data distribution is known, data points that do not follow the known distribution can be classified as outliers. There can be multiple pre-determined classes of normal data and deviation from these classes reveals outliers. Another perspective of this above-mentioned classification concerns parametric and non-parametric methods. Statistical methods often take some parameters as input whereas methods based on distance and density do not require any input parameters [1].

A completely different technique to detect groups of outliers is based on clustering algorithms where the idea is to identify clusters that are smaller in size, which are classified as outliers [2]. Clustering algorithms such as K-means are ideally suited to find convex clusters but these algorithms are highly sensitive to input parameters and can result in misclassification of clusters and outliers [3]. Usually clustering algorithms attempt to balance clusters of varying sizes. For example, spectral clustering algorithms have gained in popularity due to their low time complexity and scalability, but they use RatioCut and Ncut to create balanced clusters [4], thus making them impractical to detect rare events. Some clustering algorithms allow the generation of different size clusters [5], but *a priori* knowledge about the size of clusters is required to detect rare events, which is difficult in most domains. Furthermore, different clustering algorithms can result in different clusters for the same dataset. With all of these described inconsistencies in clustering algorithms, it is hard to rely on them to detect true positives, especially in critical applications.

Cluster based approaches often use densities and distances to identify outliers. For instance, DBSCAN [6] is the most common density-based clustering algorithm. Its notion of density reachability allows the detection of clusters of arbitrary sizes and shapes, but it cannot handle clusters of different densities. A different approach is based on the notion of isolation [7]. These methods take advantage of the fact that anomalies occur rarely in datasets. Based on training using sub-samplings and evaluation stages, these methods discover rare events by building forests of binary trees. These methods are effective in revealing global rare events, but perform sub-optimally when rare events are close to the general behavior of the entire population. In LOCI [8], the detection of outlying clusters depends on the choice of the number of nearest neighbor MinPts that define the local neighborhood. Actually, the detection of very small clusters requires a MinPts large enough to contain all points in a cluster, i.e. larger than the size of the cluster. LOCI thus proposes a multi-granularity deviation factor (MDEF) and identifies outliers as points whose neighborhood size is significantly different than the neighborhood size of their neighbors. It then relies on an appropriate choice of the neighborhood size and requires the maximum radius of the neighborhood as input parameter. A new approach, called RARE, [9], has recently been defined and proposes a two-step process to extract rare cells. Firstly, they prune the search space by removing obvious clusters that do not contain rare events. Secondly, they carefully grow these clusters, preserving their consistency. Although this approach has proved efficient for extracting rare events, the major drawback is its dependency on the input parameters required, which are hard to pre-determine.

Another approach to detecting outliers is the use of summary statistics and visual representations. Boxplots, along with its several variations [10], have commonly been used to compare univariate distributions as well as for the detection of outliers. These visual representations are hard to read and not scalable with multivariate data.

Recent advances in visual analytics and visual data mining have also introduced approaches to detect outliers through visual representation and interactive exploration [11]. Visualization techniques exploit the human pattern recognition capacity to detect anomalies and are developed by building user interfaces and interactions to deal with the graphical representation of data [12]. The problem with these approaches is their limited application to large datasets as it becomes hard to interactively explore and find rare events in thousands and millions of data items.

Extensive literature is available in the form of surveys and books to address the outlier detection issue [2, 13, 14, 15, 16] but these works do not approach the problem of detecting rare events. In this paper, we introduce a novel method to address this issue from an interactive visualization standpoint, and demonstrate that proper visual encoding is a powerful technique to explore very large datasets.

1. N. Suri, M Murty, and G Athithan. Data mining techniques for outlier detection. Visual Analytics and Interactive Technologies: Data, Text, and Web Mining Applications, page 19, 2011.
2. V. Chandola, A. Banerjee and V. Kumar. Anomaly detection: a survey. ACM Computing Surveys 41(15), 2009.
3. S. R Gaddam, V. V. Phoha, and K. S. Balagani. K-means+ id3: A novel method for supervised anomaly detection by cascading k-means clustering and id3 decision tree learning methods. IEEE Transactions on Knowledge and Data Engineering, 19(3): 345-354, 2007.
4. U. Luxburg. A tutorial on spectral clustering. Statistics and Computing, 17(4): 395-416, 2007.
5. S. Zhu, D. Wang and T. Li. Data clustering with size constraints. Knowledge Based Systems, Elsevier 23:883-889, 2010.
6. M. Ester, H.-P. Kriegel, J. Sander and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD 1996), 1996, pp. 226–231.
7. F. Tony Liu, K. M. Ting, and Z.-H. Zhou. On detecting clustered anomalies using sciforest. In Machine Learning and Knowledge Discovery in Databases, pages 274-290. Springer, 2010.
8. S. Papadimitriou, H. Kitagawa, P. B. Gribbons and C. Faloutsos, LOCI: Fast outlier detection using the local correlation integral. Proceedings of the 19th International Conference on Data Engineering (ICDE 2003), 2003, pp. 315–326.
9. E. Székely, P. Poncelet, F. Masseglia, M. Teisseire and R. Cezar. A density-based backward approach to isolate rare events in large-scale applications. In Proceedings of the 16th International Conference on Discovery Science (DS 2013), pp. 249-264, 2013.
10. R. McGill, J. W. Tukey, and W. A. Larsen. Variations of box plots. The American Statistician, 32(1): 12-16, 1978.
11. M. C. Hao, D. A. Keim, U. Dayal, and J. Schneidewind. Business process impact visualization and anomaly detection. Information Visualization, 5(1): 15-27, 2006.
12. Y. Kou, C.-T. Lu, S. Sirwongwattana, and Y.-P. Huang. Survey of fraud detection techniques. In Networking, sensing and control, 2004 IEEE international conference on, volume 2, pages 749-754. IEEE, 2004.
13. V. Barnett and T. Lewis. Outliers in statistical data. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, Chichester: Wiley, 1984, 2nd ed., 1, 1984.
14. V. Hodge and J. Austin. A survey of outlier detection methodologies. Artificial Intelligence Review, 22(2):85-126, 2004.
15. P. J. Rousseeuw and A. M. Leroy. Robust regression and outlier detection, volume 589. Wiley, 2005.
16. M. Agyemang, K. Barker, and R. Alhajj. A comprehensive survey of numeric and symbolic outlier mining techniques. Intelligent Data Analysis, 10(6): 521-538, 2006.

2 Data Sets and Problem Definition

In this section, we define datasets used for experimentation with the aim of detecting pathologic cells. Flow cytometry is a laser based, biophysical technology used in cell counting, sorting, biomarker detection and protein engineering. Flow cytometers of the current generation have a capacity for analyzing more than 10^5 cells per second. They measure the characteristics of single cells determined by visible and fluorescent light emissions from the markers on the cells. These labeled cells pass a laser that emits light at a specific wavelength according to the specific markers attached to the cell fluoresce. For each cell, a fluorescence intensity value is collected and stored in flow cytometry standard (FCS) data files [10]. FCS data file thus consists of multi-parametric descriptions of thousands to millions of individual cells. Analyzing and sorting subpopulations widely representing immune cells (CD4 + T lymphocytes, CD8 +, B, or NK) is a common practice [8].

Patient	Disease	Total Nodes	Pathologic Cells
P1	Intracranial aneurysm	1,895,261	25
P2	Intracranial aneurysm	2,524,916	15
P3	Cancer	2,470,042	7

Table 1: Datasets used for experimentation with available ground truth.

Recently biomedical science and clinical research have addressed the problem of extracting rare events from these data files [10] with 1 × 102 to 1 × 103 cells per milliliter (ml) of blood cells for 20 ml of blood. FCS data files are then used in the monitoring of vascular diseases, oncology and infectious diseases. Rare events in these cells often occur at a very low frequency, with researchers citing this number as between 0.1% to 0.00001% of the total population [3]. Recent advances in flow cytometry have enabled it to emerge as an important tool in the systems biological approach to theoretical and clinical research [10] (See [3] for further applications of detecting rare events in flow cytometry).

Methods for analyzing FCS consist of grouping individual cell data records into discrete populations on the basis of similarities in light scattering and fluorescence [12]. This is usually done by sequential manual partitioning ('gating') by plotting different combinations of descriptors two at a time in a 2D scatter plot and then select subgroups of cells using gates. Cells within the gates are selected for further analysis and plotted in another 2D scatter plot with a different axis. The main problem with this approach is that it is tedious and can also miss potential subgroup of rare cells [4], and there are difficulties in effectively analyzing high-dimensional data [8].

3 Proposed Method

The proposed method takes tabular data as input, where each row corresponds to a data item (a blood cell) along with its numeric attributes and outputs the cluster of interest, which is a set of rare events, as shown in figure 1. The processing steps are explained below.

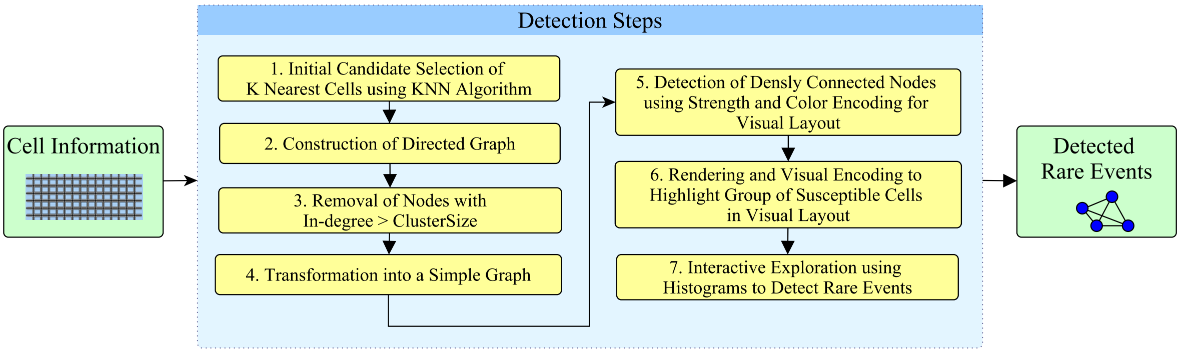

Detection Steps
Cell Information
1. Initial Candidate Selection of K Nearest Cells using KNN Algorithm
2. Construction of Directed Graph
3. Removal of Nodes with In-degree > ClusterSize
4. Transformation into a Simple Graph
5. Detection of Densly Connected Nodes using Strength and Color Encoding for Visual Layout
6. Rendering and Visual Encoding to Highlight Group of Susceptible Cells in Visual Layout
7. Interactive Exploration using Histograms to Detect Rare Events
Detected Rare Events

Figure 1: Steps of the proposed method to detect rare events.

For each pair of data items in the table, we calculate the Euclidean distance as the similarity metric among data items. The next step is to construct a graph based on this similarity among data items. For each node, we find the nearest neighbors in terms of distances using the K nearest neighbor algorithm (kNN) [5], where K is *a priori* known. For every node, directed edges are introduced where the target nodes are a node's K nearest neighbors in terms of the Euclidean distance calculated above. The choice of K value depends upon the estimated size of the rare events that we are trying to detect. For problems where this size cannot be estimated, we can interactively test different values to find an appropriate threshold. In case of pathologic cells, since we know that the usual cluster size is 50, we consider K=100 so as to ensure that we do not miss any true positives.

Hence, we obtain a directed graph where each node is connected to their K most similar data items. Since the maximum number of data items in a cluster is *a priori* known, we apply a filter to remove all nodes with an in-degree greater than the known cluster size. This is because all these nodes, which are similar to many nodes, represent regular data items that can be found readily in the graph and thus cannot belong to the group of rare events. After filtering nodes with in-degree higher than the cluster size, we ignore the orientation of edges to obtain a simple graph. This is because the edge direction is no longer used in further processing. Figure 2 shows a graph obtained as a result of the above.

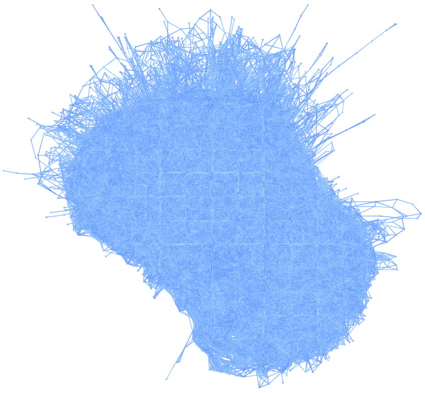


Figure 2: Visual representation of the simple graph generated in Step 4 of the proposed method. A force directed algorithm FM3 [6] (see [7] for a comparison of fast graph drawing algorithms) is used to render the graph.

The next step aims at finding clusters of nodes based on their structural similarity. We use the strength metric to detect densely connected groups of nodes within the graph. The strength metric was introduced by Auber *et al*. [2] to quantify the neighborhood cohesion of a given edge and thus determine if an edge is an intra-community or an inter-community edge within a network. It assigns nodes and edges a high value if they are connected densely to each other just like a clustering coefficient but it considers cycles of size 4 as well.

The *strength* of an edge e given by $w_s(e)$ is defined as follows:

$$w_s(e) = \frac{\gamma_{3,4}(e)}{\gamma_{max}(e)}$$

where $\gamma_{3,4}(e)$ is the number of cycles of size 3 or 4 the edge e belongs to, and $\gamma_{max}(e)$ is the maximum possible number of such cycles. Based on this definition, we define the *strength* of a vertex as follows:

$$w_s(u) = \frac{\sum_{e \in adj(u)} w_s(e)}{deg\ (u)}$$

where $adj(u)$ is the set of edges adjacent to u and $deg(u)$ is the degree of vertex u. The time complexity to calculate the strength metric over all vertices (V) and edges (E) is $O(|E|\cdot(deg_{max})^2)$, where deg_{max} is the maximum degree of the graph but in our case, since the maximum degree is bounded by a constant factor, the calculation remains constant in linear time in terms of number of edges in the graph.

Figure 3 shows the result of calculating the strength metric and applying color encoding on nodes and vertices of the graph. The scale depends on the node value interval. The tulip plugin for color mapping [1] is used. The user defines a list of colors. The first one is given to nodes having the highest values, the last one to nodes having the lowest one. A linear interpolation between consecutive colors of the list is used for the other values. In our example, nodes and edges in *red* highlight the potential rare events in the figure (Figure 3 shows the result and Figure 4 shows the color scale).

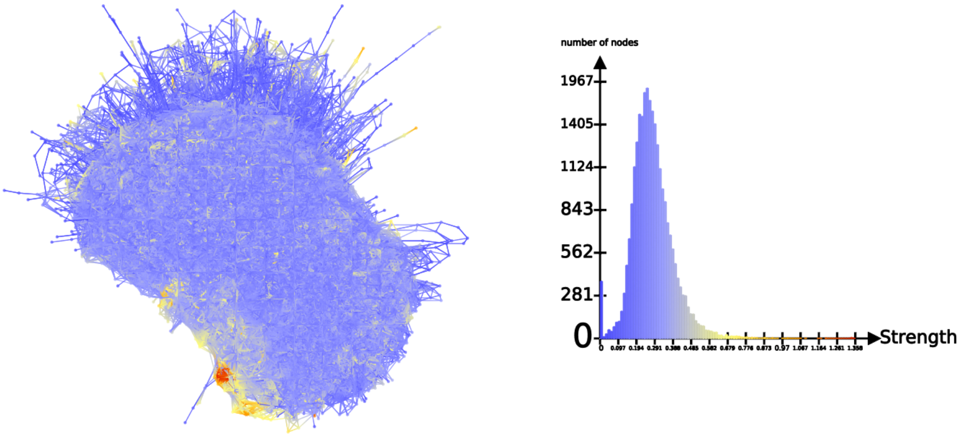
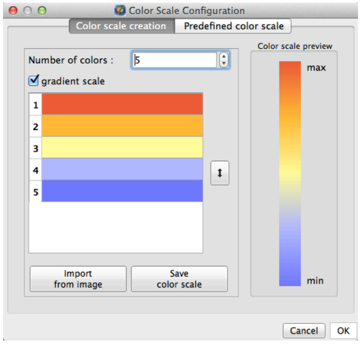


Figure 3: Computing strength metric [2] in step 6 and visual encoding it on nodes/edges, from *blue* (low values) to *red* (high values). The histogram shows the frequency distribution of different strength values. Rare events appear in red at the bottom left of the graph.


Color Scale Configuration

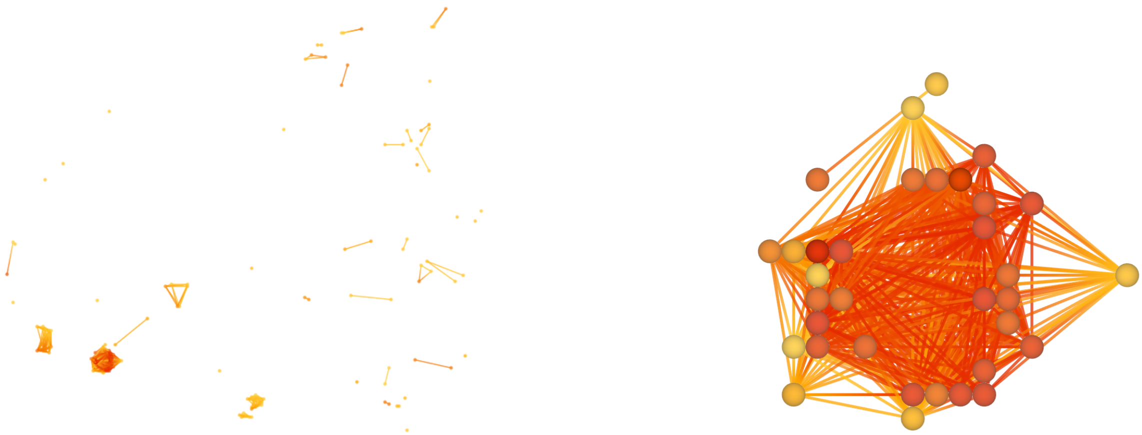
Number of colors : 5

☑ gradient scale

Figure 4: Color mapping plugin of Tulip [1]: the user defines a list of colors. The first one is given to nodes having the highest values, the last one to nodes having the lowest one. A linear interpolation between consecutive colors of the list is used for the other values.

We prefer the strength metric to the clustering coefficient because triads are more frequently present in these datasets as compared to cliques of size 4. If we use a clustering coefficient as a metric to detect densely connected groups of nodes, a very large number of true negatives will be detected. This will ultimately slow down the interactive detection process because domain experts would require further manual verification. We did not use metrics to calculate cliques of size 5 and above because such a calculation would become too slow for large datasets, as discussed in detail in the literature [9].

The final step is to visually detect the presence of rare events that form a cluster in the graph. We plot the histogram of strength of nodes along with their frequencies (as shown in figure 3) which immediately reveals that many nodes have very low strength value. Then domain experts interactively remove nodes having low strength values from the graph using histograms of frequency distribution, eventually leaving only a few nodes with high strength value, as shown in figure 5(a).



(a) (b)

Figure 5: (a) Interactive and manual removal of nodes in step 7, selecting nodes with low strength values in the histogram. Nodes with high strength could not be removed using histograms. (b) Resulting set of rare events after step 7 of the proposed method.

Domain experts then manually remove true negatives and obtain the required rare events, which is a set of densely connected nodes having high similarity to each other, as shown in figure 5(b). High similarity of nodes is depicted with color encoding, where the *red* color refers to high similarity and a gradual degradation to *blue*, which refers to low similarity among pairs of nodes.

4 Case Studies and Prototype

The proposed method has been implemented in Tulip [1]. A comparative study of different graph drawing software packages clearly shows that tulip scales well to render graphs and networks with hundreds of thousands of nodes and edges [11], thus making it the ideal platform to implement the proposed method. The attached video shows how the software helps to detect rare events on a benchmark dataset. The computation time of the strength metric clearly takes only a few seconds (even on the larger datasets described hereafter). Manual removal is also fast enough to preserve the interactivity of the tool.

4.1 Real-world datasets

A domain expert performed experiments on real-world datasets with our assistance to explain to him how to interact with the tool. He validated the results obtained using classical methods (see sidebar focusing on "Flow Cytometry").

Figures 6, 7 and 8 show the results obtained for the three datasets presented in Table 1. The results for the three datasets, when compared to the available ground truth, clearly demonstrate that the proposed method successfully found the pathologic cells within the dataset provided by doctors. In all three cases, the domain expert along with the ground truth also identified 3 to 7 false positives, which is an acceptable result since none of the true positives were missed by the proposed method.

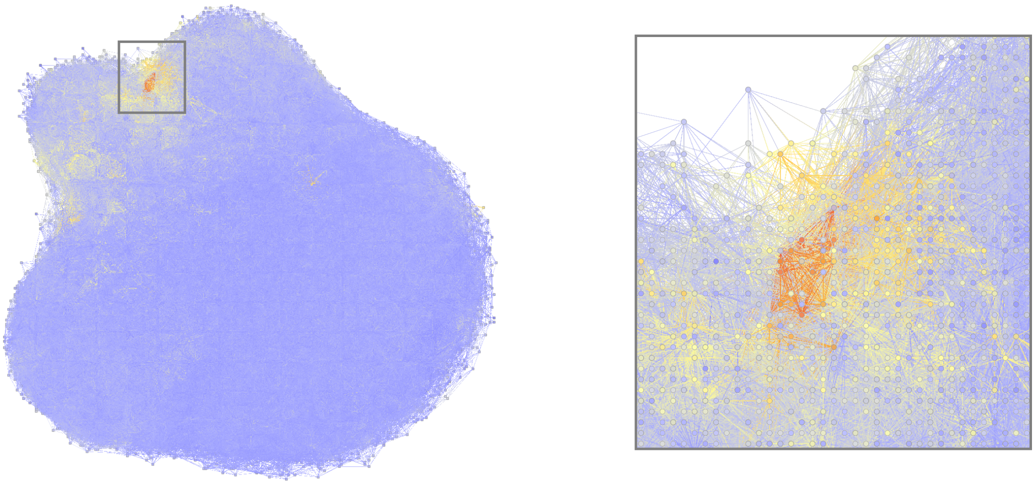


Figure 6: Patient 1: The highlighted region shows the pathologic cells found using the proposed method.

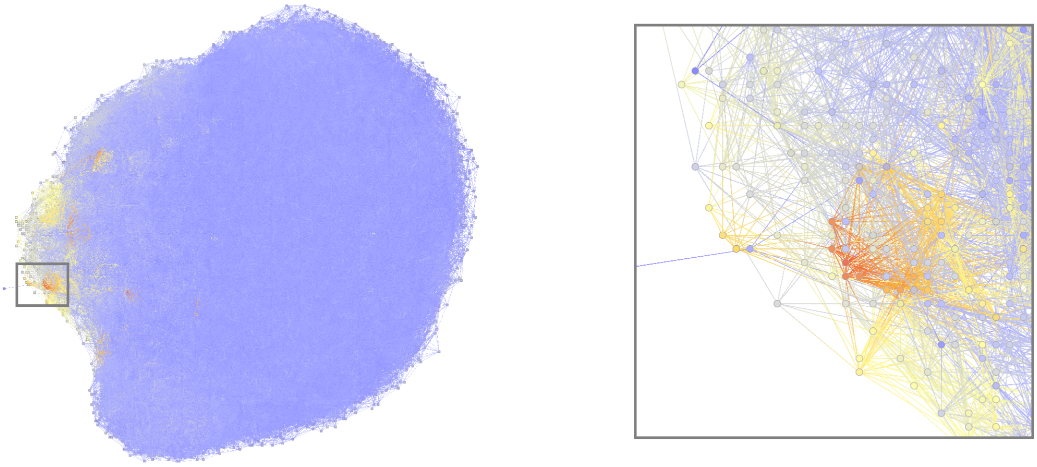


Figure 7: Patient 2: The highlighted region shows the pathologic cells found by using the proposed method.

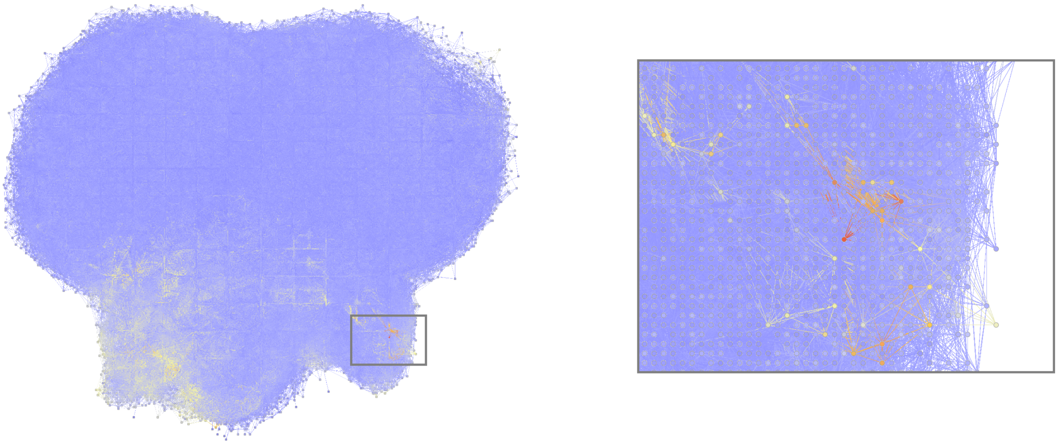


Figure 8: Patient 3: Highlighted region shows the pathologic cells found using the proposed method.

Table 2 shows the number of nodes filtered at different stages of the process as the proposed method demonstrates very high accuracy. The number of false positives detected is negligible as compared to the size of the dataset provided.

Patient	Total nodes	Nodes after Step 3	Nodes after Step 6	After removal of disconnected nodes	Pathologic cells	False positives
P1	1,895,261	126,049	33	31	25	6
P2	2,524,916	138,647	80	22	15	7
P3	2,470,042	182,626	20	10	7	3

Table 2: Number of nodes after different filtration steps of the detection process.

4.2 Benchmark

We also ran experiments on synthetic datasets. We constructed them by injecting grown pathological blood cells into a cell sample from a healthy patient. The size of the rare population injected was 50 in a dataset containing 700,000 cells.

By applying RARE on this dataset (see the "Outlier Detection Approaches" sidebar for an introduction to the methods mentioned in this paragraph), we detected 31 rare cells. Experiments were also conducted with LOCI, which is considered to be one of the best efficient approached for detecting rare events. To evaluate the best parameter setting, various values of the maximum radius in LOCI were chosen {3000, 4000, 5000, 6000}. Every time we obtained a score of 1 for points in the rare event, indicating inliers and the rare events cannot be detected. For values of the radius > 6000 we ran into memory problems. Additional experiments were carried out with DBSCAN, which reports usually high recall (generally 100%) but for most parameter values, the rare events are left unclustered and belong to the subset classified as noise.

Figures 3 and 5 show the results obtained by applying our method to the same synthetic dataset. We detected 37 cells. Unlike the real-world dataset, some pathologic cells were not detected, but all cells detected were from the rare injected population. In conclusion, even though our method does not find all the injected cells, it outperforms RARE by finding 6 more cells.

5 Conclusions

In this paper, we have proposed a novel method to detect a group of rare events in large networks. The method is based on directed networks obtained using KNN for filtering large datasets, the strength metric on edges to detect close community structures and visual encoding of the strength metric to detect densely connected groups of rare events. Results obtained on a real world biological dataset show the performance of the proposed method and clearly exhibit the accuracy superiority of the proposed method. Furthermore, the algorithm is highly efficient in terms of time complexity once we have calculated the K nearest neighbors. We intend to explore this method on other real world datasets, notably social networks, where detection of groups of outliers and rare events has many applications.

6 Acknowledgements

We would like to thank Renaud Cezar from the Institute for Research in Biotherapy, Montpellier, France. His expertise helped us in understanding the data and validating the final results. This work was mainly funded by Labex Numev.

7 References

1. D. Auber. Tulip - a huge graph visualization framework. In Petra Mutzel and Mickael Jünger, editors, Graph Drawing Software, Mathematics and Visualization Series. Springer Verlag, 2003.

2. D. Auber, Y. Chiricota, F. Jourdan, and G. Melançon. Multiscale visualization of small world networks. In Proceedings of the IEEE Symposium on Information Visualization (InfoVis), pages 75–81, 2003.

3. A. D. Donnenberg and V. S. Donnenberg. Rare-event analysis in flow cytometry. Clinics in laboratory medicine, 27(3):627–652, 2007.

4. A. D. Donnenberg, V. S. Donnenberg, and Gillian Byrne. Rapid data handling in flow cytometric rare event analysis. Biotech International, 21(1), 2009.

5. E. Fix and J.L. Hodges. Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

6. S. Hachul and M. Jünger. Drawing large graphs with a potential-field-based multilevel algorithm. In Proceedings of the 12th International Symposium on Graph Drawing (GD 2004), volume 3383 of Lecture Notes in Computer Science, pages 285–295. Springer, 2005.

7. S. Hachul and M. Jünger. An experimental comparison of fast algorithms for drawing general large graphs. In Proceedings of the 13th International Symposium on Graph Drawing (GD 2005), volume 3843 of Lecture Notes in Computer Science, pages 235–250. Springer, 2006.

8. E. Lugli, M. Roederer, and A. Cossarizza. Data analysis in flow cytometry: the future just started. Cytometry Part A, 77(7):705–713, 2010.

9. G. Melançon and A. Sallaberry. Edge metrics for visual graph analytics: A comparative study. In Proceedings of the 12th International Conference Information Visualization (IV), pages 610–615. IEEE Computer Society, 2008.

10. E. A. O'Donnell, D. N. Ernst, and R. Hingorani. Multiparameter flow cytometry: Advances in high-resolution analysis. Immune network, 13(2):43–54, 2013.

11. B. Pinaud and P. Kuntz. GVSR: an on-line guide for choosing a graph visualization software. In Proceedings of the 18th International Symposium on Graph Drawing (GD 2010), volume 6502 of Lecture Notes in Computer Science, pages 400–401. Springer, 2011.

12. N. Aghaeepour et al. Critical assessment of automated flow cytometry data analysis techniques. Nature Methods 10, 228–238, 2013.

Short author bios

Enikö Székely is a Postdoctoral researcher at CIMS, New York University, USA. She has a PhD from University of Geneva, Switzerland and her research interests include Data Mining, Machine Learning and Visualization. She can be reached at *eniko.szekely@nyu.edu.*

Arnaud Sallaberry is an Assistant Professor at the University of Montpellier 3, France. His research interests revolve around graph visualization, visual data mining, interactive visualizations and network analysis. He has been a postdoctoral researcher at the University of California at Davis and holds a PhD in Computer Science from University of Bordeaux 1, France. He can be reached at *arnaud.sallaberry@lirmm.fr*.

Faraz Zaidi is currently a postdoctoral researcher at the University of Lausanne, Switzerland. He holds a permanent position as an Assistant Professor at the Karachi Institute of Economics and Technology, Karachi, Pakistan. His research interests include data mining, information visualization, social network analysis, graphs and algorithms. He has a PhD in Computer Science from the University of Bordeaux 1, France. He can be reached at *faraz@pafkiet.edu.pk*.

Pascal Poncelet is a Full Professor at the University of Montpellier 2, France and Head of the data mining research group at the LIRMM Laboratory. His research interests can be summarized as advanced data analysis techniques for emerging applications. He is currently interested in various data mining techniques and in the definition of new algorithms for mining patterns. He can be reached at *pascal.poncelet@lirmm.fr*.

Complete Contact Information

Enikö Székely
Mailing address: Center for Atmosphere Ocean Science
Courant Institute of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012
USA
Phone: (212) 998-3224
Email: eniko.szekely@nyu.edu

Arnaud Sallaberry
Mailing address: LIRMM UMR 5506 - CC 477
161, rue Ada
34095 Montpellier Cedex 5
France
Phone: +33 467 418 653
Email: arnaud.sallaberry@lirmm.fr

Faraz Zaidi
Mailing address: Gépolis bureau 3613,
Faculté des géosciences et de l'environnement,
Université de Lausanne,
CH-1015 Lausanne,
Switzerland
Phone: +41 21 692 3076
Email: faraz@pafkiet.edu.pk

Pascal Poncelet
Mailing address: LIRMM UMR 5506 - CC 477
161, rue Ada
34095 Montpellier Cedex 5
France
Phone: +33 467 418 653
Email: pascal.poncelet@lirmm.fr