Statistically-sound Knowledge Discovery from Data: Challenges and Directions

Matteo Riondato
Amherst College
Amherst, MA, USA
mriondato@amherst.edu

***Abstract*—We describe *Statistically-sound Knowledge Discovery from Data* (StatKDD), a groundbreaking change of paradigm that shifts the focus of the KDD pipeline from the (overzealous) analysis of the available data towards understanding the, partially unknown, random, Data Generating Process (DGP) that produces the data. This shift is required by the practice of scientific research and by many industrial application, where results from data analysis *must* capture *new* knowledge about the DGP, while avoiding costly *false discoveries*.**

StatKDD considers every result obtained from the data as a *hypothesis*, which must pass *severe statistical testing* under a strong null model, in order to be considered significant, i.e., informative about the DGP.

The challenges to be solved to enable StatKDD, include *(1)* developing representative null models and severe tests for different KDD tasks from different kind of data; *(2)* considering *multiple hypotheses testing* as a necessity, not an afterthought; *(3)* offering flexible statistical guarantees, depending on the stage of the discovery process; and *(4)* creating algorithms for the extraction and testing of hypotheses that scale along multiple axes, including but not limited to the size of the data, and the number and complexity of the hypotheses.

***Index Terms*—Graph Mining, Hypothesis Testing, Markov Chain Monte Carlo Methods, Null Models, Pattern Mining, Randomized Algorithms**

I. Introduction

The Knowledge Discovery from Data (KDD) (a.k.a. Data Mining) community has developed, over the years, a corpus of ingenious methods for many analytics tasks (e.g., from pattern mining to anomaly detection) on very different data (from tabular data to graphs, to multi-variate time series), considering both static and time-evolving datasets. These approaches find widespread use in companies and in scientific labs, for applications ranging from logistics to cybersecurity, to analysis of satellite data, sports analytics, video games, and genomics research [1–11].

The KDD process (KDD) is an established pipeline in both research and production environments [12]. The process is usually described as having three steps [13, Sect. 1.2]: data collection, preprocessing, and analytical processing, feeding one into the next, with the output of processing being given to the analyst. Feedback loops from the last stage to the previous ones are considered *optional*. This design has well served the research community and the practitioners, but we believe that it fails to capture important aspects that are key to successful data-driven decision making. Incorporating these aspects is at the core of our blue sky idea of *statistically-sound KDD* (statKDD).

The KDD process, as traditionally presented, does not explicitly consider the fact that the goal of the discovery is not to better understand the available data, but rather to *gain knowledge of the process generating the data*. This idea should be evident when considering how scientific research works in, e.g., physics: the goal of performing an experiment to produce data, and then analyzing this data, is not to understand the data in itself, but rather to understand some *partially unknown mechanism* or aspect of the physical world, whose behavior or law is *captured in the data*. Understanding the data is therefore a mean to the end of *gaining knowledge about the mechanism that generated the collected data*, i.e., about the *Data Generating Process* (DGP). Thus, in the statKDD pipeline (Fig. 1), the DGP comes as the first component, and it creates the *observed dataset* (second component) which is a noisy, partial, random representation of the DGP. Knowledge of the DGP must be extracted from the dataset *while taking into account the randomness intrinsic in the DGP*.

A second aspect that must be considered is the fact that no data analysis happens in a void: the observed dataset does not get dropped from the sky to the analyst, who is unaware of how it was collected and/or of what it may represent. Rather, there exists *existing or assumed knowledge about the DGP*, which drives the analysis. The purpose of the analysis must therefore not just to extract knowledge about the DGP, but to *gain new knowledge*, "new" being the key aspect. The existing knowledge must therefore taken into account during the analysis process, which is done by defining a *null model*, which is essentially a mathematical formalization of what is currently known, or assumed, about the DGP (a more formal definition of null model is given in later sections).

The observed dataset and the null model inform the successive steps of data preprocessing and analytical processing, which are almost the same as in the KDD process. The output of analytical processing is composed of patterns, anomalies, clusters, graph/edge/vertex properties, and is considered the final result of the KDD pipeline. In statKDD, instead, these extracted quantities are considered as *hypotheses*: intermediate results that *must undergo a rigorous statistical assessment* performed with the established tools of statistical hypothesis testing [14]. Thus, analytical processing is only a "hypothesis generation stage", followed (logically, but not necessarily

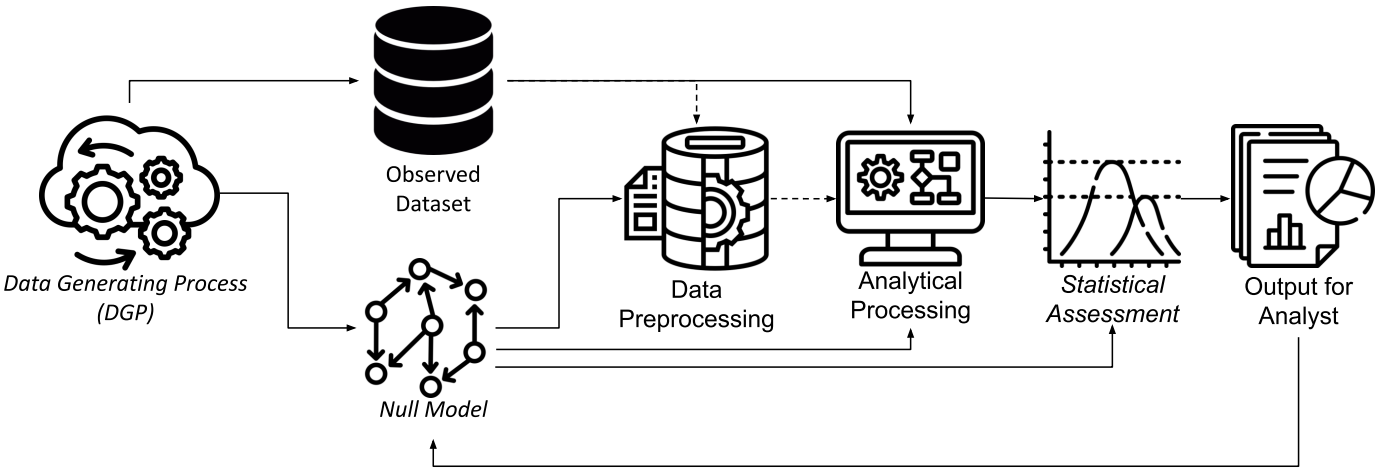

Fig. 1. The StatKDD process. *Italics* labels denote to components/steps not present in the classic KDD process. Solid connectors (vs. dashed ones) are not present in the classic KDD process.

temporarily, as they may happen concurrently) by *statistical validation with respect to the null model*: the goal of this assessment is to mark as *(statistically) significant* those hypothesis for which there is sufficient evidence, in the observed dataset, that they present *new knowledge about the DGP*, and to discard the others as due to the randomness of the DGP or not offering new information about it. The statistical validation ensures that the output of the statkDD process has *statistical guarantees* in a formal sense, therefore solving an important limitation of the existing KDD approach, where results can only be seen as giving information about the observed dataset, rather than about the DGP.

The final, but fundamental difference between statKDD and traditional KDD is the *necessity* of a *feedback loop incorporate the new knowledge about the DGP into the null model*, to ensure that future "executions" of the loop will only extract further new knowledge. Such refinement of the null model allows each iteration of the statKDD process to be informative, avoiding time spent in filtering out obvious or known results about the DGP. It is, again, perfectly in line with the practice of scientific research: when performing an experiment and analyzing the obtained data, a physicist is certainly not interested in having to filter out results that confirm Newton's law of gravitation or Galileo's isochronism of the pendulum.

We are not the first one to remark the limitations of the traditional KDD process [15], but there has been very limited work on addressing this issue, so far. In the following sections, we outline the challenges to be solved, and we present possible directions to tackle them.

II. Challenges

The following challenges must be addressed to make statKDD possible. They offer a great opportunity for researchers in different areas, and with either theoretical and/or empirical emphasis, to make a lasting impact.

1. Severe testing: It must be hard for hypotheses to be deemed significant, i.e., the statistical assessment must be *severe* [16]. The significance of hypotheses is assessed with respect to a *null model*, i.e., a collection of *possible* datasets that the (partially) unknown GDP *may* generate, and a probability distribution over this collection. The null model captures the assumed or existing knowledge about the DGP: the hypotheses are assessed against it to understand whether they can be explained by the existing knowledge or assumptions. The choice of the null model by the user must be *deliberate and informed*, as the meaning of *"statistically significant"* depends on the null model. While "all models are wrong, but some are useful" (George E. P. Box), some null models may be more appropriate for testing the significance of the results of a KDD task than others, because they *more closely represent the settings of the task*. Severe testing requires *representative null models*, thus the first challenge involves clearly analyzing the requirements and settings of data analytics task, gathering the existing knowledge about the DGP, and expressing this combined information about the DGP and the task through *(1)* constraints to the collection of datasets that can be generated, and *(2)* appropriate choices for the probability distribution over this collection. Additionally, severe testing requires that the quantities used to perform the tests, i.e., the test statistics or the (empirical) p-values associated to each hypothesis, are *conservative*, to avoid wrongly marking such hypotheses as significant: i.e., we consider more acceptable the risk (but we still aim to minimize it) of not marking a true hypothesis as significant, rather than the risk of making a false discovery.

2. Testing multiple hypotheses: The output of most KDD tasks is composed by a large number of quantities (e.g., all the interesting patterns, a score for each vertex in a graph, all the anomalies). Additionally, and key for statKDD methods to find use, the practice of science today *requires testing multiple hypotheses*: scientists do not come up with a single promising hypothesis to be tested on a well-crafted experiment that produces "perfect" data. A *family $\mathcal{H}$ of hypotheses* is

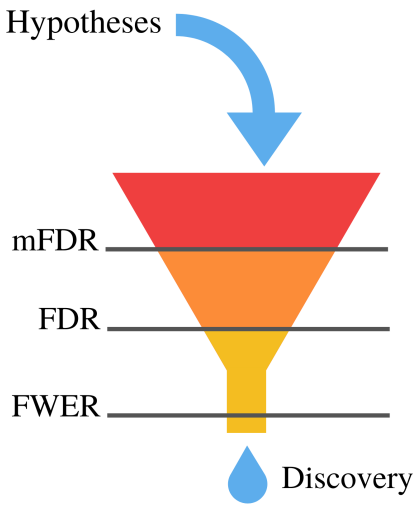

Fig. 2. The discovery funnel / still, with the stages of hypothesis testing, which act as filters for false discoveries, and the measures for false discovery control. Figure from [17].

considered, which *may* contain one hypothesis that *may* explain the phenomenon under study. For example, no molecular biologist would ever test the single hypothesis that one specific combination of gene mutations is much more often present in individuals with some disease than in healthy individuals. They would instead ask whether *any* combination of gene mutations is significantly more frequent among individuals with the disease than healthy individuals, thus testing one hypothesis per combination of mutations. The process of scientific research is then akin to a multi-stage distillation process, or to a *funnel with intermediate filters* (Fig. 2): the entire *family $\mathcal{H}$ of hypotheses* is "poured" into the funnel, and the intermediate filters, which represent different *stages of hypothesis testing* (discussed below), prevent unpromising hypotheses, i.e., those deemed to be non-significant on the observed dataset, from proceeding further. Any hypothesis that "drips" out of the funnel is a *discovery*. At each filtering stage, hypotheses are *tested simultaneously on the same data*, highlighting the need for a *multiple-hypothesis first* approach to testing, rather than as an afterthought.

3. Offering flexible statistical guarantees: Passing a single, even severe test is not sufficient to declare a discovery [16]: it only gives preliminary evidence that the hypothesis is worth further investigation. Thus, multiple stages of hypothesis testing are *necessary*. There are two *competing goals* in designing the different stages: *(1) minimizing false discoveries*, i.e., hypothesis that are false, but appear significant on the observed dataset due to the randomness in the GDP and in the testing procedures; and *(2) maximizing statistical power*, i.e., the probability to mark a true hypothesis as significant. It is extremely easy to avoid any false discovery (resp. to ensure all true hypotheses are marked as significant): one just has to avoid marking *any* hypothesis as significant (resp. has to mark *all* hypotheses as significant), but such a procedure would incur in zero statistical power (resp. would maximize the number of false discovery). There is a trade-off space to be explored in order to balance these two goals, and the trade-off point may depend on "how deep into the funnel" (Fig. 2) we are: in earlier stages, tilting towards increased statistical power is likely the right choice, while at the later stages, and definitively at the last one, it is imperative to minimize the probability of false discoveries, as we are in a now-or-never situation. Many measures to quantify the *control on false discovery* are presented in the statistics literature, e.g., the Family-Wise Error Rate (FWER) [18], the False Discovery Rate (FDR) [19], and the marginalized FDR (mFDR) [20]. Achieving a desired level of statistical power is often harder, and is usually done after ensuring that false discoveries are controlled as desired. StatKDD methods must offer *flexible guarantees* by controlling these measures, in order to be applicable at every stage of the discovery process.

4. Scaling along multiple axes: The data mining research community (and, more generally, the computer science research community) has long considered scalability with respect to input (i.e., dataset) size central to its work. But there are other axes along which statKDD methods *must* scale well, such as the *number* and *complexity* of the hypotheses to be tested.For example, the Human Protein Reference Database [21] protein-protein interaction network has ≈ 19000 proteins and ≈ 37000 interactions between them, and scientists are interested in understanding the significance of relatively small connected subgraphs in this network, representing pathways in cancer cells. There are more than 10^{13} subgraphs of size 8, each corresponding to an hypothesis. It is imperative that statistically-sound KDD methods can extract such a large number of patterns and test the corresponding hypotheses as fast as possible. In addition to *computational* scalability, statKDD algorithms must offer *statistical scalability*, i.e., perform well w.r.t. the statistical properties of false discovery control and statistical power. Figure 2 shows how hypotheses that arrive at each filtering stage must be *tested simultaneously on the same data*. Most procedures for multiple hypotheses testing consider each hypothesis in isolation, which incurs in computational and statistical "slowdowns": they unnecessarily repeat parts of the computation for each hypothesis, which limits the scalability and throughput of the filter stage, and by testing each hypothesis individually, they fail to *leverage the structure (broadly defined) of the family of hypotheses under test*, which leads to lower statistical power, thus fewer discoveries. StatKDD algorithms should *leverage* this structure, to *scale well along computational and statistical axes*.

III. Directions

Previous work towards taking into consideration the DGP show promise [22, 23], but mostly failed to tackle the challenges: it only considers simple null models, and performed tests using approximate test statistics or empirical p-values that are not necessarily conservative; in terms of false discovery guarantees, it mostly focused on controlling the FWER, which, as we discussed, is really only desirable at the later stages of the discovery process, and is excessively stringent otherwise;

it was rarely scalable, relying on Markov-Chain Monte-Carlo (MCMC) methods whose mixing time is not well studied. Finally, previous approaches were mostly designed for simple data and tasks (although, we admin, extremely important ones) such as binary transactional datasets for itemset mining, or static graphs. We propose, to the research community, the following directions to solve the challenges outlined in Sect. II.

- Propose *realistic null models for different KDD tasks* (1st challenge), informed by the needs of practitioners from different fields [24]. Good places to start are well-established tasks: various forms of pattern mining [25–28], edge/vertex centrality measures [29–31], and graph structural properties such as subgraph counts, core decomposition, and clustering coefficients [32]. Another promising task is the statistically-sound identification of anomalies, e.g., in network traffic, where anomalies may correspond to security breaches or attacks. Additionally, it is important to prove *impossibility results* showing that imposing specific constraints in the null model may lead to necessarily-inefficient testing procedures, or at least to some testing procedures being inefficient [33]: this kind of results would allow to understand the limitations of current approaches, and create interesting challenges that designers of novel methods would need to overcome.
- Derive *simultaneous confidence intervals for p-values*, to ensure that the quantities used for testing are *conservative* (1st challenge), so the control of false discovery is at the user-specified level even at finite sample sizes, not just asymptotically. A starting point could be the use of uniform convergence results from statistical learning theory, such as those based on *variance-aware Rademacher Averages* [34–36], which could obtain tight confidence intervals with little impact on statistical power.
- Design *methods to directly test multiple hypotheses* (2nd challenge), rather than considering the presence of multiple hypotheses as an afterthought. We strongly suggest embracing the *resampling-based approach to hypothesis testing* [37], which by design should allow to offer the desired *flexible guarantees* (3rd challenge), i.e., controlling, as desired, the FWER or the (m)FDR [38–40], by approximating the distribution of the p-values. Resampling methods also take into account the *structure of the hypothesis family*, which leads to good scalability with the family's size and complexity (4th challenge).
- Develop *efficient sampling procedures to quickly generate datasets* from the null model, as required by the resampling-based approach. Depending on the null model, these algorithms could be *fast-mixing MCMC methods* [32, 41] and *exact-sampling approaches* [42], that scale well along multiple axes (4th challenge).
- Subject the methods to a *thorough empirical evaluation* (5th challenge) by assessing their scalability along both computational and statistical axes (4th challenge). To help thorough evaluation of statistically-sound KDD methods beyond this project, we suggest the *development of artificial dataset generators* which allow the KDD researchers to plant true hypotheses in the generated data, so as to evaluate the tightness in the control of false discoveries and in the statistical power. The empirical evaluation should also include existing algorithms for established null models. For example, there are many algorithms available for uniformly sampling binary matrices with fixed row and column margins, but their relative performance as the size and density of the matrix changes are not clear.

IV. Conclusion

The move to statKDD from traditional KDD is motivated by the fact that abundance of data, a given fact in essentially every area of human activity, is always accompanied by a proliferation of questions whose answers should be found in such data. But the real the real questions, are about the Data Generation Process, not about the collected data, thus results obtained from the data without considering the DGP are not sufficient. When the answers are used to inform decisions that may impact large fractions of the population, e.g., for policymaking or to develop drugs, false discoveries are too costly to be allowed. Like Zimmermann [15] and others before us, we call the research community to action, but unlike them, we propose, if not a plan, a set of challenges to be solved, both computational and statistical, and some directions for how to solve these challenges. Methods that solve these challenges will have a large impact, enabling a faster, higher-throughout pipeline for scientific discoveries, and better use of data by companies and governments.

Acknowledgment

This work is sponsored in part by NSF awards IIS-2006765 and CAREER-2238693.

References

[1] E. Ferkingstad, L. Holden, and G. K. Sandve, "Monte Carlo null models for genomic data," *Statistical Science*, vol. 30, no. 1, pp. 59–71, 2015.

[2] R. T. Relator, A. Terada, and J. Sese, "Identifying statistically significant combinatorial markers for survival analysis," *BMC medical genomics*, vol. 11, no. 2, p. 31, 2018.

[3] J. Sese, A. Terada, Y. Saito, and K. Tsuda, "Statistically significant subgraphs for genome-wide association study," in *Statistically Sound Data Mining*, 2014, pp. 29–36.

[4] A. Terada, K. Tsuda, and J. Sese, "Fast Westfall-Young permutation procedure for combinatorial regulation discovery," in *2013 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2013, pp. 153–158.

[5] G. Hrovat, I. Fister, Jr, K. Yermak, G. Stiglic, and I. Fister, "Interestingness measure for mining sequential patterns in sports," *Journal of Intelligent & Fuzzy Systems*, vol. 29, no. 5, pp. 1981–1994, 2015.

[6] N. Méger, C. Rigotti, and C. Pothier, "Swap randomization of bases of sequences for mining satellite image times series," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2015, pp. 190–205.

[7] A. Mrzic, P. Meysman, W. Bittremieux, P. Moris, B. Cule, B. Goethals, and K. Laukens, "Grasping frequent subgraph mining for bioinformatics applications," *BioData Mining*, vol. 11, no. 20, 2018.

[8] T. K. Saha, A. Katebi, W. Dhifli, and M. Al Hasan, "Discovery of functional motifs from the interface region of oligomeric proteins using frequent subgraph mining," *TCBB*, vol. 16, no. 5, pp. 1537–1549, 2019.

[9] I. Alobaidi, J. Leopold, and A. Allami, "The use of frequent subgraph mining to develop a recommender system for playing real-time strategy games," in *ICDM*, 2019, pp. 146–160.

[10] C. Fan, L. Zeng, Y. Ding, M. Chen, Y. Sun, and Z. Liu, "Learning to identify high betweenness centrality nodes from scratch," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. ACM, nov 2019, pp. 559–568.

[11] T. Henderson, "Frequent subgraph analysis and its software engineering applications," Ph.D. dissertation, Case Western Reserve University, 2017.

[12] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, no. 11, pp. 27–34, 1996.

[13] C. C. Aggarwal, *Data mining: the textbook*. Springer, 2015.

[14] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, 4th ed. Springer, 2022.

[15] A. Zimmermann, "The data problem in data mining," *SIGKDD Explor.*, vol. 16, no. 2, pp. 38–45, 2014.

[16] D. G. Mayo, *Statistical inference as severe testing*. Cambridge University Press, 2018.

[17] M. Riondato, "Statistically-sound knowledge discovery from data," in *Proceedings of the 2023 SIAM International Conference on Data Mining*, 2023, pp. 949–952.

[18] C. E. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilità," *Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, vol. 8, pp. 3–62, 1936.

[19] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[20] D. P. Foster and R. A. Stine, "α-investing: A procedure for sequential control of expected false discoveries," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 70, no. 2, pp. 429–444, 2008.

[21] S. e. a. Peri, "Development of human protein reference database as an initial platform for approaching systems biology in humans," *Genome Research*, vol. 13, no. 10, pp. 2363–2371, Oct 2003.

[22] W. Hämäläinen and G. I. Webb, "A tutorial on statistically sound pattern discovery," *Data Mining and Knowledge Discovery*, vol. 33, no. 2, pp. 325–377, 2019.

[23] L. Pellegrina, M. Riondato, and F. Vandin, "Hypothesis testing and statistically-sound pattern mining," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19. New York, NY, USA: ACM, 2019, pp. 3215–3216.

[24] M. Abuissa, A. Lee, and M. Riondato, "ROhAN: Row-order agnostic null models for statistically-sound knowledge discovery," *Data Mining and Knowledge Discovery*, vol. 37, no. 4, pp. 1692–1718, 2023.

[25] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. 20th Int. Conf. Very Large Data Bases*, ser. VLDB '94. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 487–499.

[26] ——, "Mining sequential patterns," in *Proceedings of the Eleventh International Conference on Data Engineering,*, ser. ICDE'95. IEEE, 1995, pp. 3–14.

[27] M. Atzmueller, "Subgroup discovery," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 1, pp. 35–49, 2015.

[28] T. Truong-Chi and P. Fournier-Viger, "A survey of high utility sequential pattern mining," in *High-Utility Pattern Mining*. Springer, 2019, pp. 97–129.

[29] M. E. J. Newman, *Networks – An Introduction*. Oxford University Press, 2010.

[30] P. Boldi and S. Vigna, "Axioms for centrality," *Internet Mathematics*, vol. 10, no. 3-4, pp. 222–262, 2014.

[31] F. Bonchi, G. De Francisci Morales, and M. Riondato, "Centrality measures on big graphs: Exact, approximated, and distributed algorithms," in *Proceedings of the 25th international conference companion on world wide web*, 2016, pp. 1017–1020.

[32] G. Preti, G. De Francisci Morales, and M. Riondato, "ALICE and the caterpillar: A more descriptive null models for assessing data mining results," in *Proceedings of the 22nd IEEE International Conference on Data Mining*, 2022, pp. 418–427.

[33] ——, "An impossibility result for Markov Chain Monte Carlo sampling from micro-canonical bipartite graph ensemble," 2023, under submission.

[34] V. Koltchinskii, "Rademacher penalties and structural risk minimization," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1902–1914, Jul. 2001.

[35] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 463–482, 2002.

[36] C. Cousins and M. Riondato, "Sharp uniform

convergence bounds through empirical centralization," in *Advances in Neural Information Processing Systems*, vol. 33. , 2020, pp. 15 123–15 132.

[37] P. H. Westfall and S. S. Young, *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons, 1993.

[38] D. Yekutieli and Y. Benjamini, "Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics," *Journal of Statistical Planning and Inference*, vol. 82, no. 1-2, pp. 171–196, 1999.

[39] S. Dudoit, H. N. Gilbert, and M. J. van der Laan, "Resampling-based empirical bayes multiple testing procedures for controlling generalized tail probability and expected value error rates: Focus on the false discovery rate and simulation study," *Biometrical Journal*, vol. 50, no. 5, pp. 716–744, 2008.

[40] S. Denkowska, "Controlling the effect of multiple testing in big data," *Mathematical Economics*, vol. 10, no. 17, pp. 5–16, 2014.

[41] D. A. Levin and Y. Peres, *Markov chains and mixing times*, 2nd ed. American Mathematical Soc., 2017.

[42] S. Jenkins, S. Walzer-Goldfeld, and M. Riondato, "SPEck: mining statistically-significant sequential patterns efficiently with exact sampling," *Data Mining and Knowledge Discovery*, vol. 36, no. 4, pp. 1575—1599, 2022.