Assignment 4: The Last Bits

The **deadline** is July 27th at 23:59 Saarbrücken standard-time. You are free to hand in earlier. You will have to choose **one** topic from the list below, read the articles, and hand in a **report** that critically discusses this material and answers the assignment questions. Reports should summarise the key aspects, but more importantly, should include original and critical thought that show you have acquired a meta level understanding of the topic – plain summaries will *not* suffice. All sources you use should be appropriately referenced, any text you quote should be clearly identified as such. The expected length of a report is between 3 to 5 pages, but there is no limit.

For the topic of your assignment, choose one of the following:

1. *Storylines*

   Read both [1] and [2]. These papers consider the problem of making sense of large collections of text documents. Analyse both, and discuss. Are the goals clearly defined? Do the authors make principled, or ad-hoc, choices to define the solution? Are the results convincing? Are the algorithms sufficiently evaluated? What extra experiments would you have liked to see? Can you identify possible improvements for the algorithms, how would you approach the problem?

2. *Oddities*

   A large part of data mining is focused on discovering regularities in data, while oftentimes the exceptions to these underlying regularities are as, or perhaps even more informative to the domain expert. Detecting anomalies, or outliers, in graphs is particularly challenging as unlike for a row in a table—where we only regard that row and that's it—to determine whether a node (or an edge) in a graph is odd, we may have to consider *the whole graph*.

   Read two out of the following three papers, [3], [4], [5]. (Bonus, for all three) Each of these take a different look at graphs and therewith can identify different types of anomalies. Or, do they? For each method, carefully examine what makes an anomaly an anomaly, and critically discuss whether this definition makes sense, and what (hidden) assumptions these methods make. Do we always need the different methods, or can you find cases where method A find more-or-less the same anomalies as method B? Last, but not least, does there exist an ultimate outlier detection algorithm? Why (not)?

3. *Hype, hyper, hyperbolae*

   A lot of research attention in graph mining is focused on the discovery of *communities*, groups of nodes that strongly interact. Many papers assume highly simplistic models, such as that communities strongly interact within, but only loosely to the outside, or, that communities do not overlap. Recently, different groups of researchers started considering that, just like graphs themselves, communities within these graphs may also show powerlaw-like edge distributions.

   Read [6], [7], and critically discuss these approaches. Example questions you can address include the following. What are their (hidden or not) assumptions, what are the strengths, and what are the weaknesses? Consider also the evaluation, what are these methods trying to reconstruct? Is that a meaningful goal, or is it a self-fulfilling prophecy? Does any of these papers convince you that real communities follow the assumptions the authors make? The first paper considers overlapping communities, whereas the other does not. That sounds like an important downside. Is it? Really? Why (not)? (Bonus, also consider [8])

4. *Fairly Causal*

   Whenever we apply data-driven methods we have to be very careful that we do not introduce any algorithmic bias; we have to prevent our algorithms from making unfair or discriminative decisions, regardless of whether this is because of how what assumptions the method makes, or whether these biases might have been present in the data we trained the method on. Statistical parity, which assumes that all is fair if we make sure that the distribution

of the sensitive attribute is the same between the groups we decide over, is a popular notion in fair machine learning, but has been attacked by Dwork et al. [9]. Bonchi et al. [10] do not mention parity at all in their paper, as they take a causality based approach. Read this paper carefully and critically. Do you find their approach convincing? Does it live up to the critiques that Dwork gives against statistical parity? Do you see potential other problems in this approach? How could it possibly go wrong? What are the main strengths, in your opinion, and what are the main weaknesses?

(Bonus) In many applications it is not so much about making a (binary) decision, but rather about ranking elements. Say, a search engine that we can query to discover who are the smartest students on campus—and of course we want to be fair with regard to hair colour. How would you solve the fair-ranking problem, if your only tool is statistical parity? And, how would you solve the same problem if you would have the SBCN?

Return the assignment by email to tada-staff@mpi-inf.mpg.de by **27 July, 2359 hours**. The subject of the email must start with [TADA]. The assignment must be returned as a PDF and it must contain your name, matriculation number, and e-mail address together with the exact topic of the assignment.

Grading will take into account both Hardness of questions, as well as whether you answer the Bonus questions.

References

You will need a username and password to access the papers outside the MPI network. Contact the lecturer if you don't know the username or password.

[1] Shahaf, D. & Guestrin, C. Connecting the dots between news articles (../papers/a4/shahaf10dots.pdf). In Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Washington, DC, pages 623-632, 2010.

[2] Shahaf, D., Guestrin, C. & Horvitz, E. Metro maps of science (../papers/a4/shahaf12maps.pdf). In Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Beijing, China, pages 1122-1130, 2012.

[3] Akoglu, L., McGlohon, M. & Faloutsos, C. oddball: Spotting Anomalies in Weighted Graphs (../papers/a4/akoglu10oddball.pdf). In Proceedings of Advances in Knowledge Discovery and Data Mining, 14th Pacific-Asia Conference (PAKDD'10), pages 410-421, 2010.

[4] Perozzi, B. & Akoglu, L. Scalable Anomaly Ranking of Attributed Neighborhoods (../papers/a4/perozzi16scalable.pdf). In Proceedings of the SIAM International Conference on Data Mining (SDM'16), pages 207-215, 2016.

[5] Shin, K., Eliassi-Rad, T. & Faloutsos, C. CoreScope: Graph Mining Using k-Core Analysis - Patterns, Anomalies and Algorithms (../papers/a4/shin16corescope.pdf). In Proceedings of the IEEE 16th International Conference on Data Mining (ICDM'16), pages 469-478, 2016.

[6] Yang, J. & Leskovec, J. Overlapping Communities Explain Core-Periphery Organization of Networks (../papers/a4/yang14overlap.pdf). Proceedings of the IEEE, 102(12), IEEE, 2014.

[7] Araujo, M., Günnemann, S., Mateos, G. & Faloutsos, C. Beyond Blocks: Hyperbolic Community Detection (../papers/a4/araujo14hycom.pdf). In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Data (ECMLPKDD'14), Springer, 2014.

[8] Metzler, S., Günnemann, S. & Miettinen, P. Hyperbolae are no Hyperbole: Modelling Communities that are not Cliques (../papers/a4/metzler16hyperbolae.pdf). In Proceedings of the IEEE International Conference on Data Mining (ICDM'16), IEEE, 2016.

[9] Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. Fairness through Awareness (../papers/a4/dwork11fairnessawareness.pdf). In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS), ACM, 2012.

[10] Bonchi, F., Hajian, S., Mushra, B. & Ramazotti, D. Exposing the probabilistic causal structure of discrimination (../papers/a4/bonchi15causaldiscrimination.pdf). Data Science and Analytics, Springer, 2017.

(http://mmci.uni-saarland.de)