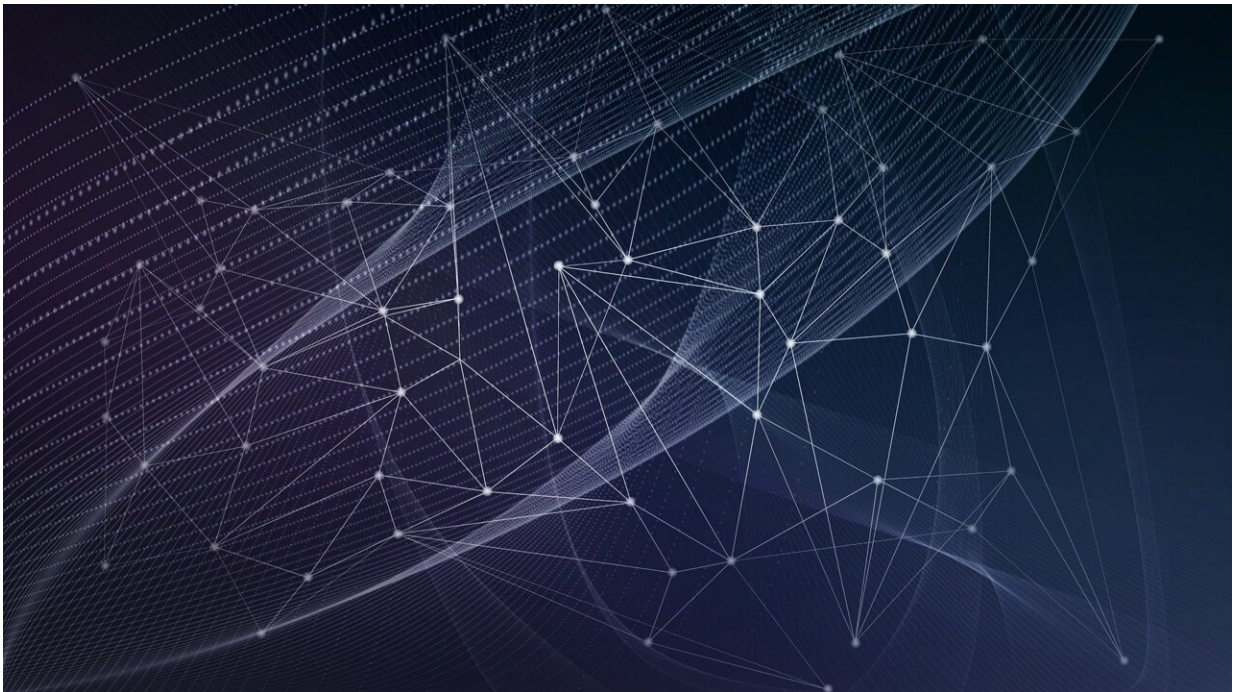# Study showcases new method for better grouping in data analysis

July 22 2024

Credit: Pixabay/CC0 Public Domain

Researchers at Carnegie Mellon University and UC Berkeley have developed a new method to improve how computers organize and analyze large datasets. This advancement improves the ability to extract information from knowledge graphs, impacting the ability to analyze social networks and customer behavior.

The new method explained in a study led by Benjamin Moseley, Carnegie Bosch Associate Professor of Operations Research at the Tepper School of Business at Carnegie Mellon, can group similar items together more effectively while keeping different items apart.

The paper will appear in ICALP, the International Colloquium on Automata, Languages, and Programming conference, which took place in July 2024.

"Our new algorithm can significantly enhance how we analyze large data sets, whether it's for improving social media platforms by accurately detecting user communities or advancing medical research by better understanding genetic interactions," Moseley said.

He noted that a key trend in business analytics is the ability to work with knowledge graphs, which show information like customer behavior or business processes. This paper focuses on clustering, a common method for extracting information from these graphs. The new method in this study can group similar items more effectively while keeping different items apart.

Organizing massive amounts of data correctly is challenging due to inconsistencies and the sheer volume of information. Moseley and his team focused on creating an algorithm that can quickly and accurately group data points. They used mathematical structures consisting of nodes, which represent data points, and edges, which are connections between nodes. The algorithm works by evaluating these connections and determining the best way to group similar nodes.

The results showed that their algorithm is faster and more accurate than previous methods. It can handle large data sets more efficiently, making it practical for real-world applications.

"Our new method is faster than any previous methods at minimizing mistakes when grouping data," said Sami Davies, a research scientist in theoretical computer science at the University of California, Berkeley. "Our method is also more flexible, in the sense that we can group data in a way that is good for many different objectives simultaneously."

The researchers plan to continue refining their method and exploring its applications in different fields. This ongoing work could lead to even more accurate and insightful data analysis.

Heather Newman, a Ph.D. candidate in the Algorithms, Combinatorics, and Optimization doctoral program at the Tepper School was also a co-author.

  More information: Sami Davies et al, Simultaneously Approximating All ℓp-norms in Correlation Clustering, *arXiv* (2023). DOI: 10.48550/arxiv.2308.01534

Provided by Tepper School of Business, Carnegie Mellon University