Mathematics and Big Data

Big Data and Topographical Data Analysis

Soh Yong Xiang, Tan Jen Han and Tan Jia Jun Shaun
Hwa Chong Institution

Abstract:

Big data exists everywhere. Its significance comes from analysis and application of the results to various aspects of society. This covers everything from government policy to social media, education, healthcare, sports and finance. One of the methods used to interpret big data is through topographical data analysis, which effectively and efficiently organises torrents of data to derive valuable information. With the information obtained, products and services can be personalised and improved. In other words, the use of topographical data analysis and big data in general has great potential for affecting positive change in our daily lives.

Essay:

Channel News Asia uses the same stock footage every morning when they talk about Wall Street. Office spaces, numbers running across black screens. Many of us have come to accept this cold, hard string of numbers as the face of banking, of data in general. Then the scene changes. We are back in the newsroom, but this time we don't know where the numbers have gone.

We ask one of our classmates from the Infocomm Club about big data, and this is what he tells us: big data is large and chaotic. While defining big data in itself may be challenging, many stick to the 3Vs: volume, velocity and variety. In other words, it's data that comes from anywhere and everywhere, that comes fast, that comes from and in a diverse range of sources and formats.

He compares big data to the server traffic that goes through Facebook. There's a large volume of data that comes in from the activity of its 1.28 billion daily active users, data that comes in rapidly from this constant worldwide activity, and data that comes in a variety of ways (browsing habits, text, video and photo uploads, personal information, location etc.). In this case, Facebook uses this information to refine the ads it shows individual users as well as to work on its features according to users' habits.

When described like that, big data almost seems personal and humane. Oftentimes, it is unwanted intimacy - there are long-running conflicts regarding data collection and privacy - but it is intimacy nonetheless. And the data collected isn't just for marketing, of course, since the significance of big data ultimately lies in its application to real-world situations at large.

Our classmate mentions hospital records. Data gathered includes patient check-ins, conditions and treatment. But the data gathered in healthcare is much more diverse than that. There are still health insurance records, wearable devices, lifestyle apps. With all this data, we can generate a profile of each individual, with potential diseases and preventive solutions. This has tremendous potential for the future of healthcare, where accurate targeting of prospective patients allows for efficient prevention, rather than cure, saving on both time and cost for constant, new treatments. In some ways the idea of intimacy returns. In return for big data's benefits, we trade the often overlooked parts of our private lives. We disclose both our past and our present as necessary parts of ourselves: anywhere from our family's genetic history to the number of calories we've eaten and steps we've taken today.

Education is similar. Rather than hard reliance on results from standardised testing, specifics from every student's learning process throughout the school day can be taken into consideration: the way we do questions, what kind of questions we cannot do, the depth of online learning done and social interaction between peers along with its relationship with academic performance. Then comes information like socioeconomic class, attendance, school demographics, information that is equally important to the learning process as larger-scale social factors. All of these are used to personalise online and offline curriculum based on students' learning profiles. Weaker students may receive more help, potentially ineffective programmes and policies may be cancelled, learning tasks may be tailored to build on students' interests. All of these are possibilities for education big data affords.

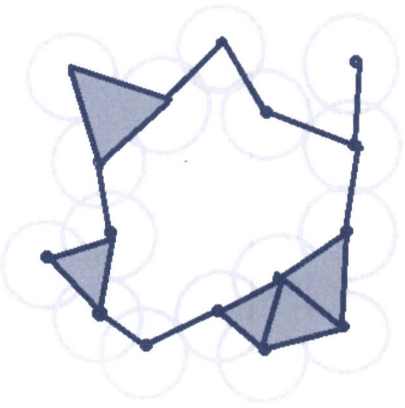
In fact, some of this is already going on, albeit on a much smaller scale. Think of teachers, pulling out weaker students for additional lessons that focus on their specific weaknesses. Amidst all the talk of great possibilities big data has for the world, it is important to remember its small, everyday applications. Applications that now seem real and relevant, even in our own small worlds.

Still, big data is called big data for a reason. What is intimacy to us is inconsequential when placed alongside the billions of others worldwide, whose endless output of data is analysed and seen besides ours to detect trends and patterns. Especially with the advent of the internet, sources of data are no longer limited to the limited resources immediate to us, but all the world's citizens at large.

So how do we understand all this data? Our classmate talks about graphs. More specifically, topological data analysis (TDA) - an approach in applied Mathematics that uses techniques from topology which enables us to link related data sets to derive information.

Topology is a wide field of research. For analysis of big data, a more specific method used is persistent homology. Persistent homology is an algebraic method for discerning topological features in data. These algebraic features include components, holes, and graphs. TDA and persistent homology are used where normal analyzing software fail, such as when data lies not in the form of a linear graph, but as a mass of data that cannot equate to a graph. Topology is powerful in analyzing these data sets, manipulating the data and removing insignificant "noise", the data can be converted into their topological forms and patterns, trends or anomalies can be identified.

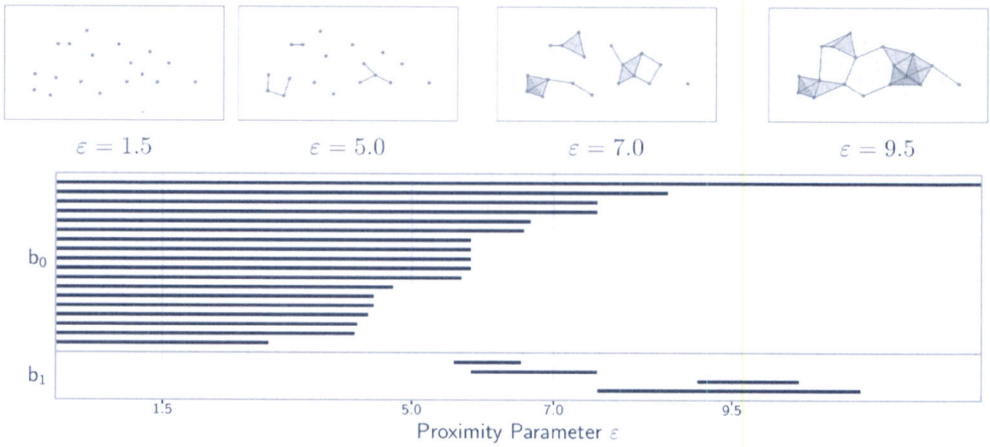
In topology, a simplicial complex is an object of connected points, edges, triangular faces etc. Persistent homology is like counting the number of links, holes and voids within a simplicial complex.
Data, especially big data, often appears as a point cloud, each point representing a measurement of some object or data. Each of these points can be visualized to have a circle of a radius of d around them, also known as its proximity parameter, in which d is a variable. As d increases, this imaginary circle increases in size, causing some to overlap. Whenever two points overlap, the points are connected by a line. As it increases further, complete simplices of multiple points may be formed as more points overlap each other. Such complete simplices are filled in with triangular faces or 3-dimensional simplices. Eventually, we get a simplicial complex known as a Rips Complex. We can apply homology to this complex, which may reveal holes or voids.



An example of a simplicial complex in Persistent Homology

However, if d is too small, circles will not overlap, creating multiple small holes or "noise" that may be irrelevant to the data. If d is too large, eventually the any two points can be connected, and a giant simplex is formed, which has insufficient information and cannot be effectively analyzed. Since d is a variable, we

can present it as a graph, also known as a bar code. As d increases, points gradually become connected, and small or large holes may begin to appear. When they appear, they are represented as a bar, which is known as that hole's persistence.. When holes are eventually filled in by edges as d grows, the bar stops. A collection of these bars effectively track the Persistence of holes across the various values of d, also known as the bar code. Existing standard computer algorithms can then be used to compute and analyze the barcode, or through linear algebra.



A conventional output from persistent homology is a "bar code" graph.

Short bars are small packets of data that appear to be irrelevant and insignificant, often arising from "noise" in data or anomalies, and may be disregarded, depending on the context. Longer bars suggests significant features in the data that appear to overlap, that can potentially be easily analyzed, or modelled to provide significant or key information.

TDA is superior to other forms of data analysis in many ways, as it can identify trends, or important features that were overlooked or previously hidden. It allows us to understand the stream of never ending, complex data without having to simplify it to view it. Furthermore, bar codes are stable. That is, when the points of data are adjusted a little, the bar code has minimal adjustments. This is important in real life applications, as measurements will always have some margin of error.

The possibilities of its application are wide, and endless. An example of this is in finances, or specifically in the stock market. In this case, TDA is able to discern patterns in massive big data sets that would otherwise be overlooked. Normal data mining is often limited computationally, while the use of persistent homology will be able to search out unknown patterns that are hidden, allowing firms to patch up pricing anomalies without delay.

Another possible application would in sports. For example, a hockey team manager could utilise TDA to find out what players would be most suitable for their team, and what teams are most likely to be strong competitors. With the endless possibilities for every team's composition, traditional methods of simply picking a player based on their apparent skill during a match is sometimes unreliable. On the other hand, with the use of Persistent Homology in TDA, bar code diagrams can be drawn up to identify the deficiencies in the team's composition. The manager can then fix these deficiencies to improve their team composition by hiring another player best suited for the team chemistry, ideally creating the "perfect" team.

Ultimately, we cannot say if big data is an intimate thing or not. It is simultaneously vast and specialised, real yet abstract. But we do know this: it has been and will continue doing great things in various areas of our lives, whether in healthcare, or education, or even sports, just to name a few. It also has powerful uses in banking - uses that are far, far beyond just the numbers on the screen during Channel Newsasia broadcasts. This, we know.

References:

1. Brandon. (2017). *Proximity Parameter* [Graph]. Δ Quantitative √ourney. Retrieved July 6, 2017 from http://outlace.com/TDApart1.html
2. Dutcher, J. (2014). *What Is Big Data?*. Retrieved July 3, 2017 from https://datascience.berkeley.edu/what-is-big-data/.
3. Goldfarb, D. (2014). *AN APPLICATION OF TOPOLOGICAL DATA ANALYSIS TO HOCKEY ANALYTICS* [PDF file]. Retrieved July 06, 2017, from https://arxiv.org/pdf/1409.7635.pdf
4. Knudson, K. (2017). *Topology looks for the patterns inside big data*. Retrieved July 6, 2017 from http://theconversation.com/topology-looks-for-the-patterns-inside-big-data-39554
5. Lafrance, A. (2014). *Facebook Is Expanding The Way It Tracks You And Your Data*. Retrieved July 3, 2017 from https://www.theatlantic.com/technology/archive/2014/06/facebook-is-expanding-the-way-it-tracks-you-and-your-data/372641/.
6. Marr, B. (n.d.). *How is Big Data Used in Practice? 10 Use Cases Everyone Must Read*. Retrieved July 3, 2017 from https://www.ap-institute.com/big-data-articles/how-is-big-data-used-in-practice-10-use-cases-everyone-should-read.
7. Marr, B. (2015). *How Big Data Is Changing Healthcare*. Retrieved July 6, 2017 from https://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/#50f4982e
8. Rabella, M. F. (2016). *How does big data impact education?*. Retrieved July 6, 2017 from http://oecdinsights.org/2016/11/07/how-does-big-data-impact-education/2873
9. SAS Institute. (n.d.). *What Is Big Data?*. Retrieved July 3, 2017 from https://www.sas.com/en_sg/insights/big-data/what-is-big-data.html.
10. Symonds, J. *Topological Data Analysis and Finance*. (2013, August 06). Retrieved July 6, 2017 from https://www.ayasdi.com/blog/topology/topological-data-analysis-and-finance/
11. van Rijemenam, M. (n.d.). *Four Ways Big Data Will Revolutionize Education*. Retrieved July 6, 2017 from https://datafloq.com/read/big-data-will-revolutionize-learning/206