Is Amazon.com the Same Everywhere?

A Regional Comparison of Product Diversity on Amazon.com in California and Missouri.

Team Members

Deirdre Murphy, Margaux Reynders, Joana Stockmeyer, Romane Donadini

Contents

Summary of Key Findings

Amazon.com does provide anonymous users with a different product selection depending on the location they set as their delivery address. In this preliminary research, no specific pattern has been found as to whether Amazon.com favors one region over another one in terms of product quality, brand selection or prices. However, each query did return different results for at least one of the four users, each of them having set a different delivery address. The display of the results for each query also differs depending on the location: different ads might be featured, different product arrangement as well. Amazon.com also shows products of different economic values to the different users. However, we did not notice any price discrimination for the same product between the different set regions.

1. Introduction

Amazon is a large international company that uses data to personalise results based on a vast variety of indicators from the user. A study suggests that 35% of the company's revenue comes from their recommendation system (Mackenzie et al.) For this reason, the company's algorithms learnt to use data to personalise results, and we began to see a shift from uniform results to personalised by default. Today, Amazon search results are the consequence of complex correlations working together based on a great amount of different variables. Also, because these algorithms are continuously changing and moving it is impossible to identify the exact reasons for individual results.
According to Gillespie, algorithms play an important role in selecting what information is most relevant to each person (167). Furthermore, search engines enable us to navigate colossal databases of information, and "recommendation algorithms" provide us with items or pieces of information that we might prefer instead of others, "suggesting new or forgotten bits of culture for us to encounter"
(167). Amazon's marketing strategy relies on these algorithms to show consumers more personalised and refined results in order to increase the possibility of a sale. But what happens when a user does not have an account, order history, or even data for the platform to track? Is there still diversity and differentiation in the shopping space between different users in different regions? Throughout this project, we wish to investigate what is coined the "cold start problem," which suggests that in order to provide best results, the system needs to have some information about the user (*medium.com*). By searching different queries with a clean browsing history, Amazon is forced to randomise results

as it cannot personalise in the early phases. We seek to examine if these randomised results differ for individual users with clean browsing histories in different regions.

2. Initial Data Sets

We used twelve different data sets. Six of them were collected with a clean search browser for anonymous users and the other six were collected with a browser whose search history had not been cleaned and as logged in users on Amazon.com. For both the clean browser method and the logged-in one, each of the six data sets corresponds to one of the six queries we made: [playstation], [candle], [mouse], [coffee], [nail polish], [lamp]. In other words, we have a data set for each query for both logged-off and logged-in accounts. All of them contain the search results of the four different anonymous or logged-in users with their corresponding delivery address. Each set has a column for the products' value, its original price, the applied discount, the pseudo of the user, the number of acquired products -or amount of displayed products for each user-, the product ID, the search ID, the query itself, the saving time, the order of appearance of the product, a thumbnail, a link to the product itself, the average amount for all the displayed products for each query.
We combined the six data sets collected with the clean method in one spreadsheet, and the six data sets collected with the logged-in method in another spreadsheet. We added a column to both of them, displaying which method was used.

3. Research Questions

To what extent is the diversity of products on Amazon.com reflected in different regions?

4. Methodology

Our methodology evolved throughout the project. Initially, we wanted to test if personalisation on Amazon affected pricing and products displayed regionally. We planned to benefit from the diversity of our group's four people, including: one American account, one French account, one German account, and one person not having an account at all. Then we would simply set the delivery address to The Netherlands. We quickly had to modify our idea as we wanted to set a standard currency (USD) and language (English), for coherence purposes. Since the three group members with diverse accounts also all owned an American account in addition to that, we decided, instead, to look at different states within the United States. We also realized that the personalisation of each of our accounts would greatly affect the results we would get for each searched product on Amazon.com. Thus, we decided to use logged-off accounts and undertake the searches on Brave, a clean search browser. By having the different delivery addresses as our only variable, we made sure that the collected data would not be influenced by the previous purchase history of our personal accounts.
We set four different delivery addresses, with four different postal codes. Two of them are located in California: 90804 (Long Beach) and 94116 (San Francisco), respectively for Southern California and Northern California. The other two are located in Missouri: 63301 (Saint Charles) and 64030 (Grandview), respectively for East and West Missouri. Each one of us searched the same queries, each at the same time and using the same spelling, namely: [playstation], [candle], [mouse], [coffee], [nail polish], [lamp]. These items were chosen as they differ highly in their price range and fall under different categories.
To be able to collect the results, we downloaded two browser extensions. The first one is called Full Screen Page Capture, with which we took a full screenshot of the result page of each query. We re-named the files obtained for organisational purposes and classified them in a specific folder. The second extension is named Amazon Tracking Exposed (amTrex) and was designed to research Amazon tracking, enabling us to download the product search results in CSV files, which we also classified in specific files and folders for more clarity.
To visualise the data, we used two different softwares: Gephi and Tableau. With Gephi, we first had to create a node sheet on Google sheets, with three columns:
"Id," "Label," and "Type." In the Id column, we entered the four different locations, the query, and the products' Ids that appeared in the search results for each person. We also deleted the duplicates, so there would be only one unique node per product. For the label column, we entered the same information as for the Id. The type column was filled with "location," "query," and "product." We then created the edge sheet, with two columns: "Source" and "Target.".
In the column source we first entered the locations and matched them to the name of the query in the target column. Then, we entered the results of the search for each query, with the location in the source column and each

product found in the search results for each location in the target column. We uploaded the edge sheet to Gephi in a new workplace, then, we uploaded the node sheet to the same workplace. We ran the “Force Atlas 2” layout, with the options to dissuade hubs and prevent overlaps. We personalised the colors of the nodes by type (purple for the products, green for the query, and orange for the locations) and changed their size, ranging from five to 13, depending on the query and the amount of nodes.
For Tableau, we imported the CVS of all clean query data to Tableau and created a new sheet. On the sheet, we dragged “product id” to a row and excluded any items that were “null” or did not have a valid “product id.” We then arranged them in alphabetical order for a more coherent visualisation. After this, we dragged the “number of records” to the top column to create the sum. We also arranged them in the most common appearance ranging from four to one. From there on, we dragged “pseudo” to “color” to distinguish which regions were shown each product, and then we dragged query over to a row and deleted “product id” from the row. Lastly, we dragged “pseudo” over as well to a row and arranged them in “sorted descending by sum of number of records within pseudo”.

5. Findings

Generated with the Tableau software, Figure 1.1 and 1.2 display the product diversity present in the four different delivery regions of North California, Southern California, East Missouri and West Missouri. The bars represent the number of products shown in each region, where Figure 1.1 presents all of the products that were in the result lists - excluding duplicates - and Figure 1.2 displays products that appeared uniquely in a certain region. Figure 2.1 and 2.2, created with the help of Gephi, show the connection of products related to the query [nail polish] that showed up for one, two, three, or all four of the regions. Finally, Figure 3 displays the average price of the products available in the search results for each query by location.

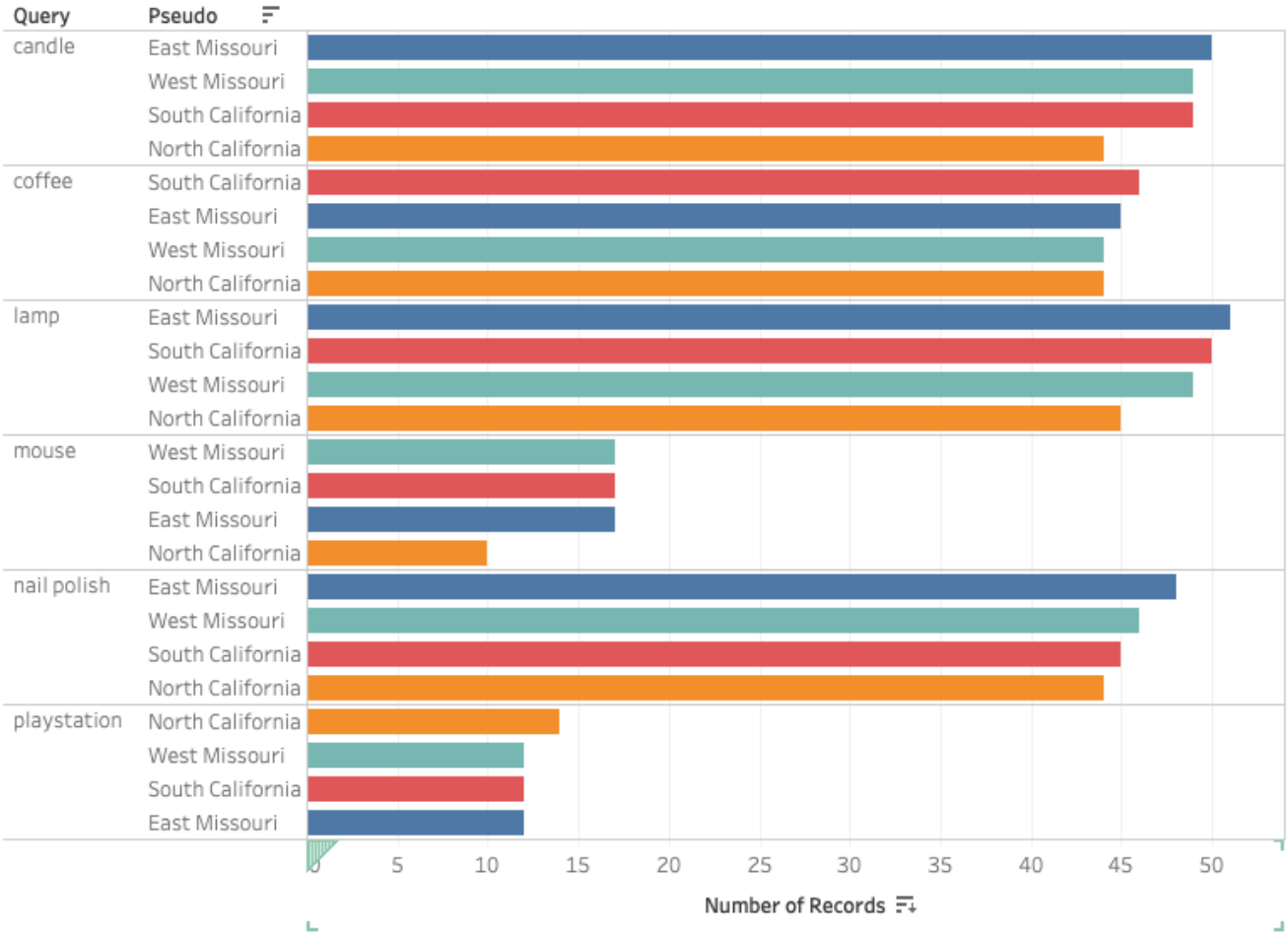

Figure 1.1 All products to appear, excluding duplicates.

As displayed in Figure 1.1, the search queries for [candle], [lamp], and [nail polish] showed similar results in regards to the general product diversity. All of these queries resulted in an approximate average of 45 to 50 results excluding any product that appeared more than once in all of the regions with North California showing the fewest results in each case. West Missouri and South California had similar numbers, while East Missouri always displayed the highest numbers. For [coffee] the numbers were marginally lower with South California presenting the highest number of singular results

(46). [Mouse] and [playstation] on the other hand had a significantly lower number of singular products. East and West Missouri as well as South California showed 17, while North California had 10 for the query [mouse]. Only for the query [playstation] North California displayed the highest number of singular results (14), while the remaining three regions all showed 12 singular products.

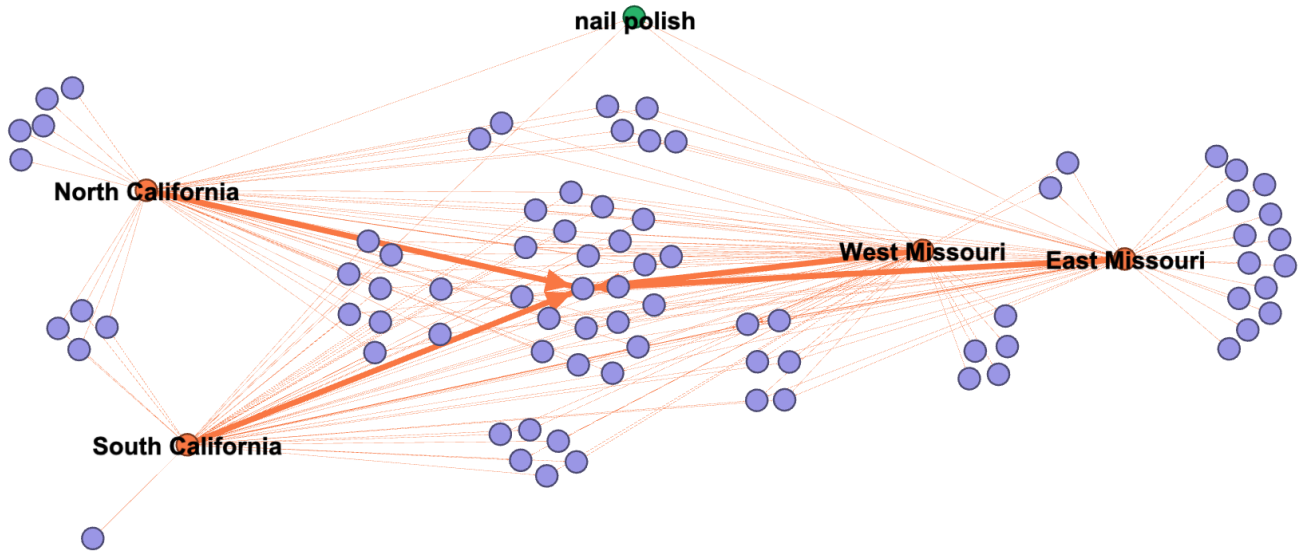

Figure 1.2 Products appear only appear to one pseudo= uniqueness

Figure 1.2 displays the uniqueness of the products returned by Amazon.com for each query and each location. For the queries [candle] and [coffee], the searches with North California as the delivery location showed the greatest amounts of unique results. On the other hand, the query [lamp] provided both West and East Missouri with the higher number of unique products (7). The query with the highest amount of unique results is [nail polish], which provided the user with East Missouri as their delivery region with 14 unique products. The results for the queries [mouse] and [playstation] featured very similar results for every location, except for West Missouri, which got 3 distinctive results for [mouse] and North California, which received 4 distinctive results for [playstation]. The only delivery location which consistently received the least amount of unique products and thus the most generic results is South California.

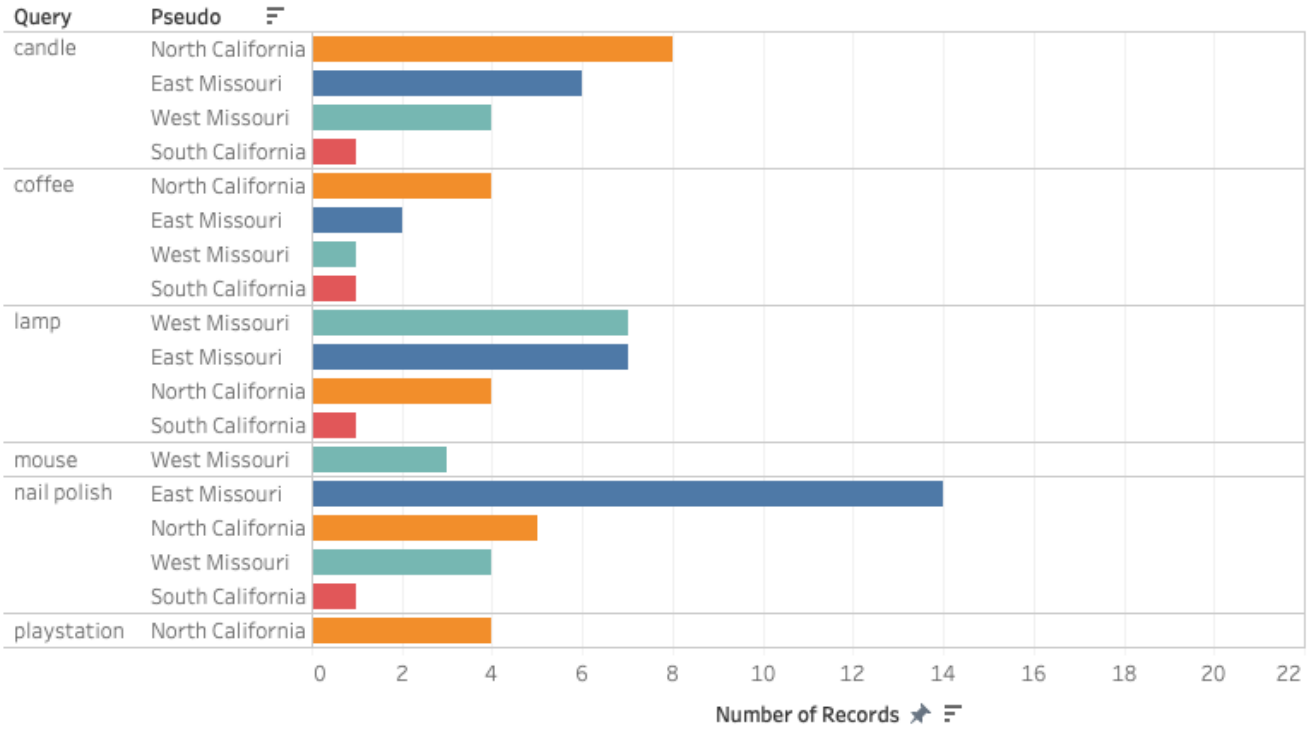

Figure 2.1 Connection of products returned for the query [nail polish] for the delivery regions of North California, South California, West Missouri and East Missouri.

Figure 2.1 is a network representation, made with gephi, of the product selection provided by Amazon.com for the query [nail polish] for the four different delivery regions. The four orange nodes illustrate our chosen delivery

locations, while each purple node represents a different product that is connected to one, two, three, or all four of the locations with orange lines depending on whether it showed up in the results list or not. We can see that the majority of the returned products are shared by the four locations. They are located in the centre of the illustration; however, they are surrounded by smaller groups of nodes connecting two or three of the locations as several products are only shown to specific locations. The fewer connections a purple node has, the more it is located towards the borders of the illustration.

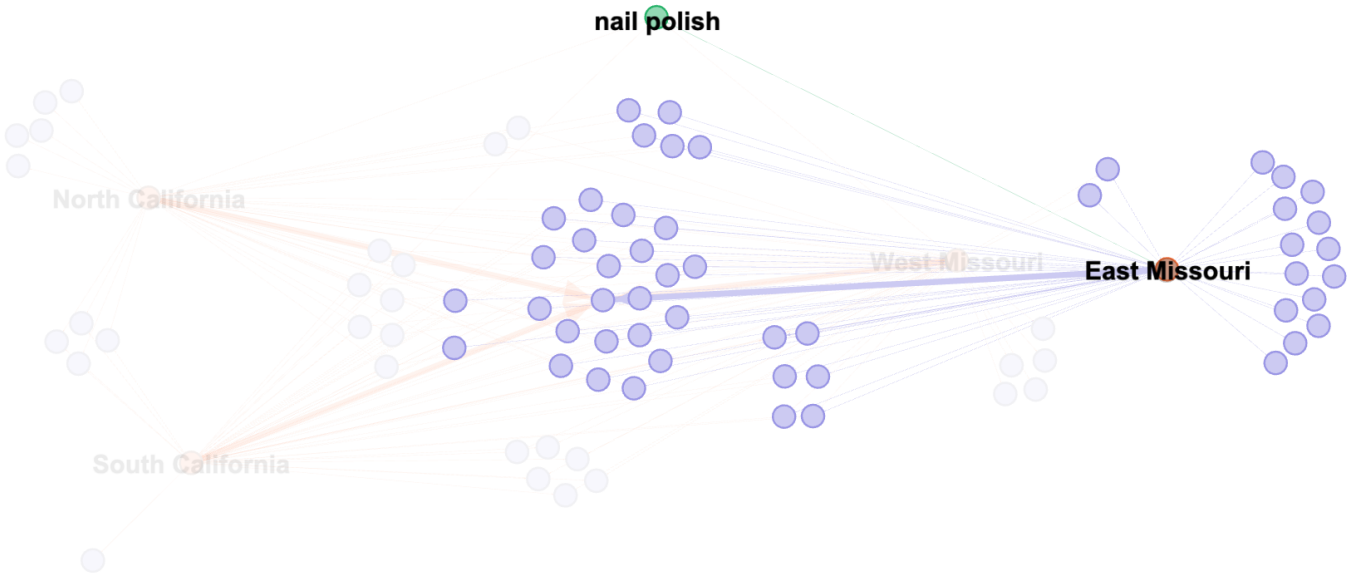

Figure 2.2 Connection of products returned for the query [nail polish] for the delivery region of East Missouri.

Figure 2.2 highlights the products displayed for the delivery location of East Missouri for the same query. It clearly shows that a large amount of the nodes are specific to East Missouri and are not shared with any other location. 14 nodes showing up on the border on the right represent the amount of unique products featured in the results of the [nail polish] query for the anonymous user whose delivery location was set to East Missouri. When comparing this focus to Figure 2.1, it becomes clear that the other locations did not receive as many distinctive product results: one for South California, five for North California, and five for West Missouri.

In addition to the diversity of the selection of products provided by Amazon.com for each query, the average price for the sum of the product selection also differs depending on the location. In two instances, the location of West Missouri showed a higher average than the other locations and together with East Missouri, it featured the highest average price of $15.6 for the query [coffee]. The query with the greatest discrepancy in the average price is [lamp], which averaged at $42.4 and $36 for Northern California and

West Missouri respectively. This is a difference of $6.4, which is quite considerable, and represents 15.09% and 17.78% of the two respective averages. The query [mouse], with a discrepancy of $0.9, is the one that provided us with the smallest price difference between the four locations.

6. Discussion

We know that Amazon's personalisation affects the results provided to logged in users as their purchase history and other personal characteristics are taken into account by the website's algorithm. Following this logic, we could think that anonymous users would receive the same results for the same queries, especially when the searches were made on a clean research browser. However, analysing the collected data allows us to demonstrate that even for anonymous users, Amazon.com returns different results for the same query depending on the person's shipping locations.

We do not have the means to explain the reasons behind the discrepancy in the product selections offered by Amazon. This being said, we can presume that several specific factors influence what we have observed. Among them, we can consider the differences in the consumer habits by region. We can imagine that Amazon might already know what kind of products users in specific locations usually purchase and thus mostly offers those same products to people located in the same locations. The average wealth or purchasing power per location might also affect the selection of products returned by Amazon for each query. Lastly, the population size and type of each shipping location might be determinant factors when it

comes to the products shown by the website. Indeed, the demographics of each region can potentially influence what is mostly bought and what is not.
The idea of the existence of the "cold start" problem for the Amazon recommendation system is still applicable, even if we did observe a slight difference between the several locations and their results. By looking at the data we collected as logged-in users, we were able to see a much greater discrepancy in the product diversity and in the price averages for each single query. Amazon.com's recommendation system does indeed work much better when it possesses some data on the user.

7. Conclusion

With the assistance of the amTREX tool, we were able to investigate to what extent is the diversity of products on Amazon.com reflected in different regions. By searching the queries: candle, playstation, lamp, mouse, coffee and nail polish, we hoped to distinguish any correlation or differentiation in the results based on region. To do this we made sure that we were all logged out of any Amazon account and cleared our browser's, Brave, data. Ultimately, we are not able to identify the specific reasons for the causation in results; however, it was clear from the data collection that there is discernment between the regions. For example, Southern California had the most generic results based on the queries we searched. We also determined that based on the query, some products proved more generic than others. For example, there are more variations of types of nail polish or lamps than there is of playstations. This means that the diversity was similar in all four regions for the query playstation.
In order to further investigate this matter, we believe that the experiment could be extended to more states in the United States for Amazon.com, but maybe also to different regions for other Amazon's shopping spaces, such as Amazon.fr, Amazon.de, Amazon.co.uk (etc.). Furthermore, by searching for more queries, we could also look at any major pattern in the way Amazon offers its products to specific regions.

8. References

'Approaching the Cold Start Problem in Recommender Systems'. Medium, 28 Apr. 2017, https://medium.com/@InDataLabs/approaching-the-cold-start-problem-in-recommender-systems-e225e0084970.
Burrell, Jenna. 2016. How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. Big Data & Society 3(1).
Gillespie, Tarleton. 2014. The Relevance of Algorithms. In: Gillespie, T, Boczkowski, P J and Foot, K A (eds.) Media Technologies: Essays on Communication, Materiality, and Society. Cambridge MA: MIT Press, 167-194.
Gitelman, L. (ed.) 2013. Raw Data Is an Oxymoron. Cambridge: the MIT Press. Introduction chapter.
Kitchin, R. 2014. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. 1 edition. Thousand Oaks, CA: SAGE Publications Ltd. Chapter 1: Conceptualising Data.
MacKenzie, Ian, et al. How Retailers Can Keep up with Consumers | McKinsey. https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers. Accessed 28 Jan. 2020.
Rieder, Bernhard, Ariadna Matamoros-Fernández, and Òscar Coromina. 2018. From ranking algorithms to 'ranking cultures'. Investigating the modulation of visibility in YouTube search results. Convergence 24(1): 50-68.
Yeung, Karen. 2017. 'Hypernudge': Big Data as a Mode of Regulation by Design. Information, Communication & Society 20(1): 118-136.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 825974-ALEX, with Stefania Milan as Principal Investigator; https://algorithms.exposed).