Advanced Social Data Analysis 1

Exam
April 2021
University of Copenhagen, MSc in Social Data Science
Exam numbers: 45, 50 and 51
Characters: 47,880 corresponding to 19,95 pages

The number in the margin indicates the student responsible for the paragraph in question.

Exam nr. 45, 50 and 51
Exam in Advanced Social Data Science 1 - 03/04/2021

Content

Part 1: Using a Regression Discontinuity Framework to estimate the effect of digital nudging at Elgiganten.dk 2
Introduction 2
Data 2
Inventory status messages 4
Analytical approach 6
Motivation for RDD 6
The actual cutoffs triggering status messages 6
The baseline model 6
Causality and bandwidth - the nonparametric approach 7
Pooling the discontinuities 8
Results 8
Covariate balance 11
Discussion 12

Part 2: Measuring ties among students at DTU through a comparison of two digital networks 15
Introduction 15
Data 16
Modifications of the datasets 16
Step 1. Comparison of the two networks 16
Social cohesion - Network density 17
Each person's social connection to cohort - Nodal degree 17
Social stratification - Group degree centralisation 18
Results 18
Step 2: Defining modes of social relatedness 20
Measures of modes of social relatedness 21
A weighted multigraph 21
Results 22
Consequences of modes of relatedness 24
Discussion 24
Conclusion 25

Literature 26

Appendix A.1 - Extracted data from Elgiganten.dk. A.1a is data scraped at 28-03-2021 and A.1b is data scraped at 29-03-2021
Appendix A.2 - Jupyter notebook for data analysis of Elgiganten.dk data as .pdf (A.2a) and as .ipynb (A.2b)
Appendix B.1 - Data from the Copenhagen Network Study. B1a - edgelist call network, B1b node for call network, B1c - edgelist for Facebook network, B1d - nodes for Facebook network
Appendix B.2 - Jupyter Notebook for network analysis as .pdf (B.2a) and as .ipynb (B.2b)

Part 1: Using a Regression Discontinuity Framework to estimate the effect of digital nudging at Elgiganten.dk

Introduction

For this part of the exam we explore the effects of digital nudging on an online platform of the electronic appliance retailer Elgiganten. Digital nudging can be defined as the use of design
45 elements in a user-interface to guide behaviour in digital choice environments (Weinmann et al. 2016:433). The nudging element of interest for the present exam is a written message visible for each product telling the customer whether there are fewer than 5, more than 5, more than 25, more than 50, or more than 100 of a product in stock. The message is a nudge implemented by Elgiganten as an incentive to increase their sales with the underlying assumption that a customer will be more inclined to buy a product if they know that there are only few in stock. Thus, our hypotheses is:

> *A message shown to a customer indicating a lower inventory status for a product increases the likelihood of a customer making a purchase of the product.*

By using data from Elgiganten's online platform, we estimate the impact of the nudge on customer behaviour. Customer behaviour is measured as a dummy variable indicating if a
50 purchase of a product has been made within 24 hours after recording which inventory status message (nudge) is associated with the given product. To make valid causal inference, we use a sharp regression discontinuity design (RDD) to ensure a quasi random distribution of customers receiving the different inventory messages. We conduct a covariate balance check to assess the assumption of local randomisation. We find that there is no causal effect of inventory messages on the likelihood of a product being purchased. However, we also discuss several concerns with both the data and the analytical framework applied.

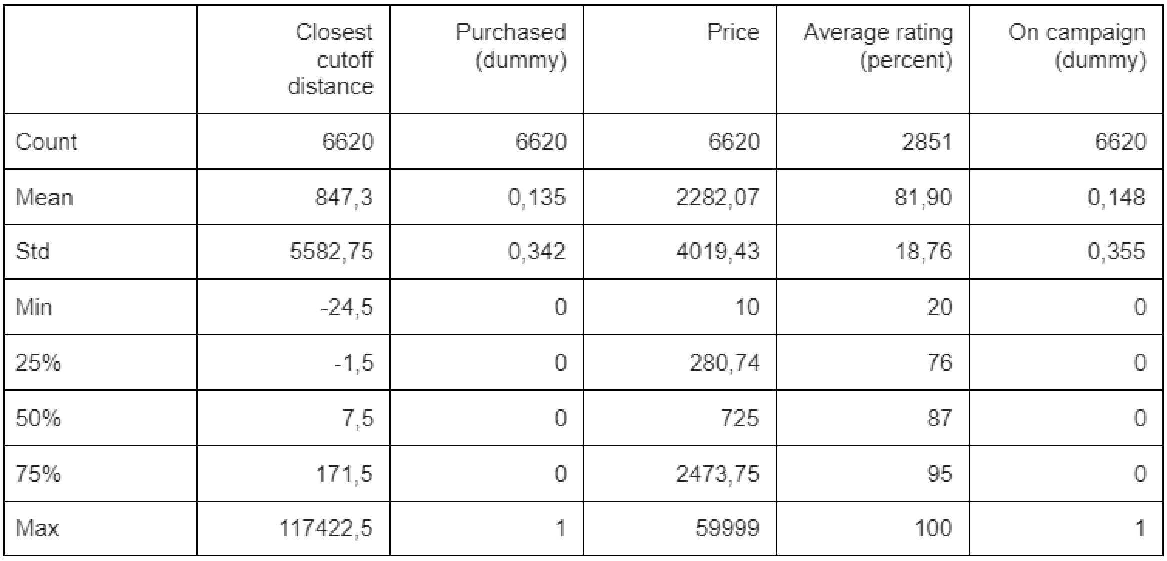
Data

In order to conduct this analysis, we have retrieved product data on 14,130 Elgiganten products by using a REST API called when navigating the website, Elgiganten.dk. The 14,130 products in
51 the dataset are all the products listed by searching for the wildcard character "*" on the

Elgiganten webpage[1]. We are able to extract the inventory level that determines which messages customers receive from the API.

From the original dataset of 14,130 products, we remove some observations. First, we remove observations (N=3,476) that do not have a listed inventory level as this information is crucial to construct variables related to which inventory status message customers receive. These are products where either Elgiganten has no set amount, examples are products such as warranties and gift cards. Second, we remove products where Elgiganten has zero inventory (N=4,034), as customers looking at those products receive an inventory status message telling them when Elgiganten expects to restock, which is another type of message than the nudges of interest in this project. We end up with a dataset containing 6,620 products in total (46.9 pct. of original dataset). Table 1 describes the descriptive statistics of some of the variables of interest in the dataset.

Table 1- Description of the data after removing products that Elgiganten do not have in inventory



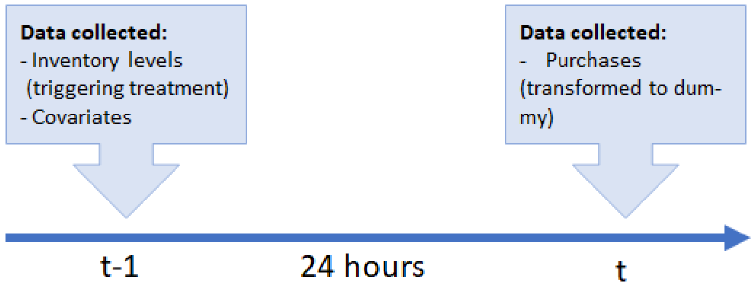
	Closest cutoff distance	Purchased (dummy)	Price	Average rating (percent)	On campaign (dummy)
Count	6620	6620	6620	2851	6620
Mean	847,3	0,135	2282,07	81,90	0,148
Std	5582,75	0,342	4019,43	18,76	0,355
Min	-24,5	0	10	20	0
25%	-1,5	0	280,74	76	0
50%	7,5	0	725	87	0
75%	171,5	0	2473,75	95	0
Max	117422,5	1	59999	100	1

Note: "Closest cut off distance" is the distance in inventory level to the closest point where inventory status message changes. "Purchased (dummy)" is a dummy for whether a product has been purchased or not. "Price" is the price of products in DKK. "Average rating (percent)" is a review measure from 20-100 (converted a 5-point rating system to numbers), not all products have received ratings hence the lower count. "On campaign (dummy)" is a dummy for whether the product is on campaign.
Source: Data in Appendix A.1 and own calculations

[1] The search results on the Elgiganten webpage using the Elgiganten search engine are limited to 10,000. However search results can be ordered in different directions (high-low price, low-high price, alphabet, product release date and 5 other ways). Using these different search result ordering we retrieve the list of approx. 14,000 products.

The data is collected at the beginning and at the end of a 24 hour period. The timing of the data collection is illustrated below in Figure 1:

Figure 1 - Timing of data collection


Data collected:
- Inventory levels (triggering treatment)
- Covariates
Data collected:
- Purchases (transformed to dummy)
t-1
24 hours
t

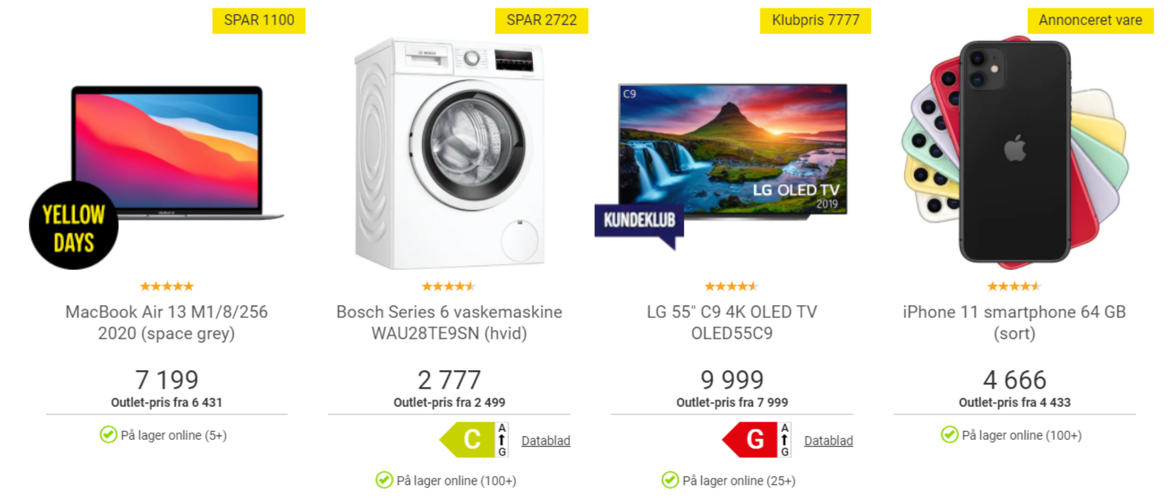
Note: “t-1” is scraped at “29-04-2020 11:00 CET” and “t” is scraped at “30-04-2020 11:00 CET”
Source: Authors’ own illustration.

As shown in Figure 1, data about all of Elgiganten’s products is first collected at time $t - 1$ and then again at time t. From the data collected at $t - 1$, we extract the inventory levels and the 50 inventory status message and different covariates such as price, campaign status, and ratings. From the data collection at time t, we extract the number of purchases of each product within the last 24 hours. This mode of data collection does, however, have certain implications: We do not have information about the inventory status message for each individual purchase. Thus, we cannot be sure that customers receive a specific inventory status message if several products with inventory levels near the thresholds, that trigger different inventory status messages, are purchased within the 24 hours outlined in Figure 1. Nevertheless, we know the inventory status message for the first purchase of each product within the timeframe. This is the reason why we define our outcome as whether a purchase of a product has been made or not instead of the number of times a product has been purchased.

Inventory status messages

When using the webpage, customers are presented with the written messages of inventory levels at two separate places. Whenever a customer navigates the front page and search 51 functions they are presented with a list of products. Here, inventory messages are presented underneath the price of each product as shown in Figure 2.

Figure 2 - Elgiganten message - on overview sites



SPAR 1100

MacBook Air 13 M1/8/256 2020 (space grey)
7 199
Outlet-pris fra 6 431
På lager online (5+)

SPAR 2722
Bosch Series 6 vaskemaskine WAU28TE9SN (hvid)
2 777
Outlet-pris fra 2 499
C Datablad
På lager online (100+)

Klubpris 7777

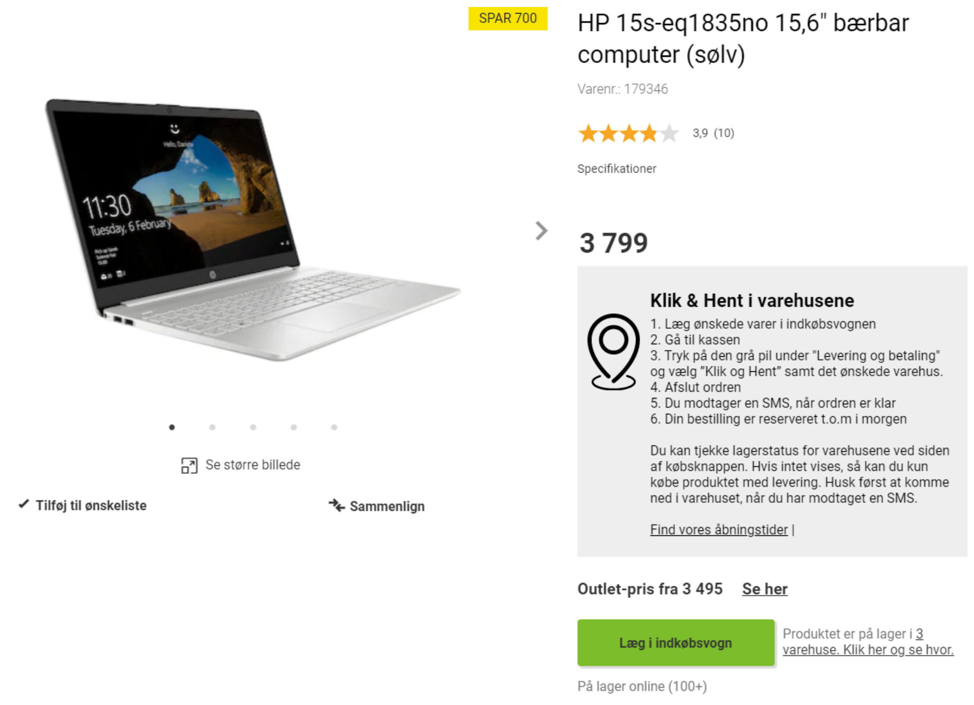
LG 55" C9 4K OLED TV OLED55C9
9 999
Outlet-pris fra 7 999
G Datablad
På lager online (25+)

Annonceret vare
iPhone 11 smartphone 64 GB (sort)
4 666
Outlet-pris fra 4 433
På lager online (100+)

Source: www.Elgiganten.dk

Second, when clicking on a product, customers are presented with a host of related information including the message on inventory levels as shown in Figure 3. There are thus two possible locations on the website, where inventory messages are visible to customers.

Figure 3- Elgiganten message -on product site



SPAR 700

HP 15s-eq1835no 15,6" bærbar computer (sølv)

Varenr.: 179346

3,9 (10)

Specifikationer

3 799

Klik & Hent i varehusene

1. Læg ønskede varer i indkøbsvognen
2. Gå til kassen
3. Tryk på den grå pil under "Levering og betaling" og vælg "Klik og Hent" samt det ønskede varehus.
4. Afslut ordren
5. Du modtager en SMS, når ordren er klar
6. Din bestilling er reserveret t.o.m i morgen

Du kan tjekke lagerstatus for varehusene ved siden af købsknappen. Hvis intet vises, så kan du kun købe produktet med levering. Husk først at komme ned i varehuset, når du har modtaget en SMS.

Find vores åbningstider |

Outlet-pris fra 3 495 Se her

Læg i indkøbsvogn

Produktet er på lager i 3 varehuse. Klik her og se hvor.

På lager online (100+)

Source: www.Elgiganten.dk/product/179346/

Analytical approach

Motivation for RDD

45 In order for us to test our hypothesis we estimate the impact of a change in inventory status message, the treatment, on the likelihood that a product is purchased, the outcome. We deploy a regression discontinuity design (RDD) which is especially useful in handling possible selection bias issues. These issues arise as the inventory levels are possibly correlated with the purchasing behaviour of customers separately from the inventory status message. As an example, Elgiganten might have relatively large quantities of a given product in stock if they expect this product to be popular, or if it is less cost-intensive to store. Thus, products with different messages might be different with regards to their inherent properties, and a comparison of these products might correspond to a comparison of apples and oranges.

50 To handle the selection bias, we use a sharp RDD and exploit the inventory level cutoffs that trigger these messages to the customers. This means that we only compare purchase behaviour of products where the inventory level is either just above or just below 5, 25, 50 and 100. This way we assume as-good-as random treatment assignment for these products. However, a concern could be that Elgiganten decides how much to keep in stock of a given product by using the same cutoffs in their inventory strategy e.g. instantly automatically restocking selected products that drop below an inventory level of 5. This could lead to different types of products just below and just above the cutoffs implying a non-random assignment of treatment. Furthermore, the interpretation of an effect of the treatment might instead be an effect of Elgiganten's inventory strategy in general. By checking whether relevant covariates balance across the cutoff we will get an indication of the possibility of the cutoffs being used in Elgiganten's inventory strategy.

The actual cutoffs triggering status messages

51 When the inventory level of a product goes from 4 to 5 the inventory status message shown to the customer changes from "<5" to "5+". As a consequence of inventory levels taking discrete values, the cutoff triggering the change in message is assumed to be 4.5 as an inventory level of respectively 4 and 5 will have the same distance to the cut off value. Applying the same logic to the triggering of the inventory status messages "25+", "50+" and "+100" the total list of possible inventory message cutoffs are 4.5, 24.5, 49.5, and 99.5.

The baseline model

Our approach to RDD is similar to that outlined by Angrist and Pischke (2009, 2015). Our baseline model is:

$$y_i = \alpha + \rho D(x_i) + \epsilon_i, \text{ where}$$

$$D(x_i) = \begin{cases} 1 & x_i \geq 0 \\ 0 & x_i < 0 \end{cases}$$

$$x_i = z_i - c_i^*$$

$$c_i^* = \arg\min_c |z_i - c|, \text{where } c \in [4.5, 24.5, 49.5, 99.5]$$

y_i is a dummy variable indicating whether product i has been purchased within the last 24 hours. $D(x_i)$ is a treatment dummy, that is a deterministic function of the running variable x_i which is the difference between the inventory of product *i*, z_i, and the closest inventory stock cutoff c_i^*. c is the cutoff that triggers a new treatment message measured 24 hours before the outcome variable. α is a constant parameter to be estimated, and ϵ_i is the idiosyncratic error term.

Causality and bandwidth - the nonparametric approach

We rely on a nonparametric estimation approach to motivate causal inference and ensure that ρ can be interpreted as a treatment effect. We impose a restriction on the data so it consists only 45 of observations within an interval around the cutoff value triggering the treatment. The interval can be defined as the distance k (where $k > 0$) to the cutoff point: $c - k < x_i < c + k$. Within distance k we assume it is as-good-as random whether products fall above or below the cutoff and thus whether treatment is assigned.

The distance $k$is referred to as the bandwidth, and the size of it can impact the results of the analysis. A large $k$implies that more observations are included in the estimation of the model, but it also implies a larger likelihood that the products differ in their inherent properties leading to a higher risk of selection bias. However, if$k$is too small, selection bias is less likely but there might be too few observations in the interval to identify treatment effects.

50 The running variable is increasing stepwise due to it being measured in integers. This imposes restrictions on the value of k, which might impact the validity of the assumption of conditional random assignment of treatment. It is especially problematic when considering an isolated treatment effect of the customer receiving the message “5+” instead of “<5” , as there are only 4 negative integer distances to the cutoff 4.5, that the running variable can take. There is a higher risk of selection bias as the outcome variable is distributed on few discrete values of the running variable.

We use an algorithm derived by Imbens and Kalyanaraman (2009) to find the optimal bandwidth. This is a data driven approach that has been proved to perform well (ibid.:17). Another benefit of using this algorithm is that it decreases researchers' degrees-of-freedom and thus makes it more difficult to manipulate k to find the most remarkable results. Nevertheless, we will consider different bandwidths from the optimal when estimating the model to examine the impact of changing bandwidths on the estimation. Furthermore, we check for covariate balance conditioning on the choice of optimal bandwidth to test the validity of the conditional random assignment of treatment assumption.

Pooling the discontinuities

51 We choose to pool the discontinuities instead of estimating the treatment effect of each discontinuity individually. Pooling the discontinuities implies that the effect of receiving the message of "<5" or "5+", "5+" or "25", "25+" or "50+" and "50+" or "100+" is combined. Hence, the estimated treatment effect can roughly be interpreted as the share of unique products purchased given that a customer is told that there are "more" products in stock compared to the share of unique products purchased if the customer is told that there are "few" products in stock. Pooling increases the statistical power as the model will include more observations close to the pooled cutoff. However, we believe that the effects might be heterogeneous between the differences in inventory status messages. As an example, the effect on customer purchase behaviour of receiving the message "<5" products in stock compared to the message "5+" products in stock might be very different to the effect on behaviour of receiving the message "50+" products in stock compared to the message "100+" products in stock. To check the validity of this assumption, we estimate the baseline model for each of the discontinuities and explore if there are heterogeneous effects on customer behaviour that complicates the choice of pooling the cutoffs.

Results

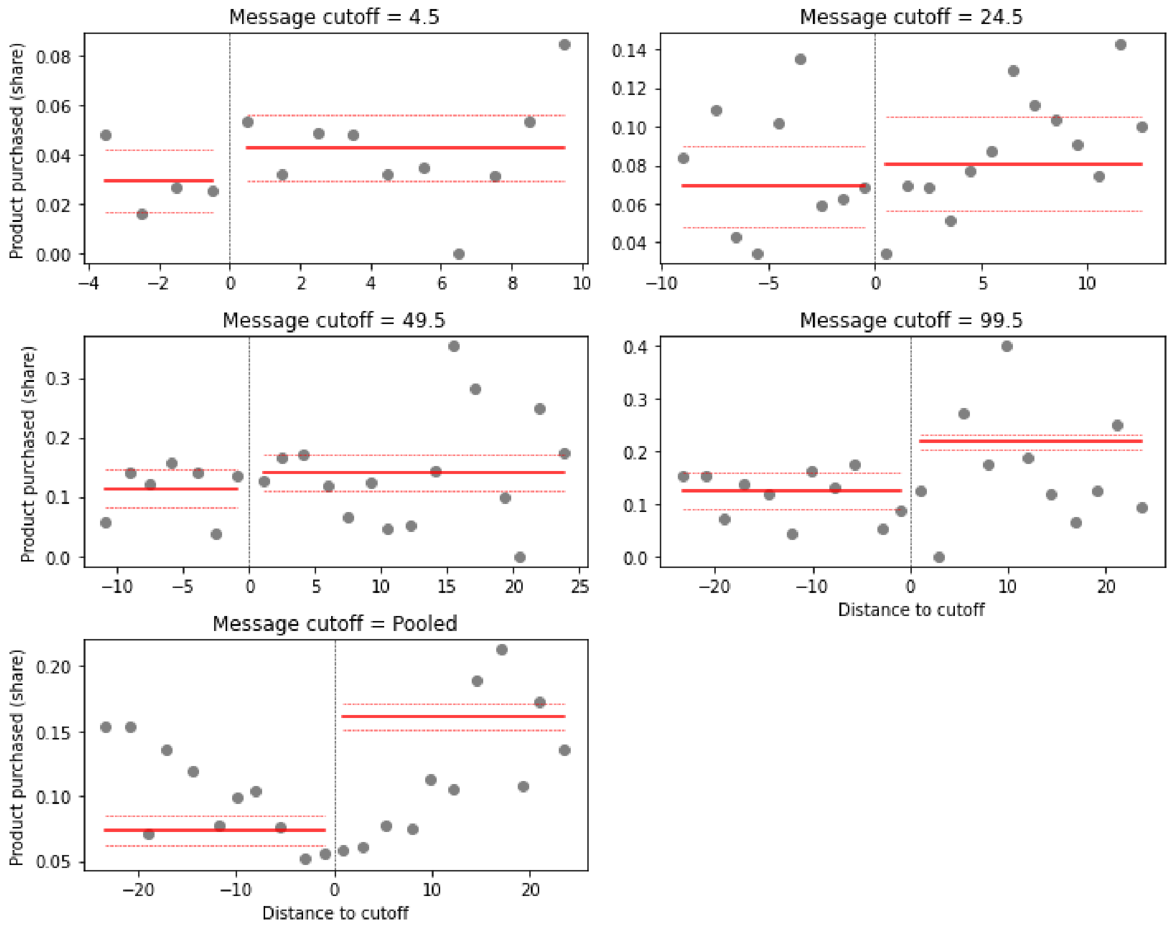
45 In this section,, we start by presenting the motivation for introducing bandwidths to the baseline model. Then, we estimate the model using bandwidths with our preferred model specification. When estimating the baseline model, we find no significant effect of the different inventory status messages on the share of unique products purchased.[2]

In Figure 4, we present our baseline model estimated without a bandwidth. This estimation is equivalent to comparing the mean of all products that have received treatment with all products that have not received treatment. In Figure 4, it is clear that there is a significant difference in the share of products purchased between products with different inventory status messages

[2] Our code and data for this section is attached in respectively Appendix A.1 and A.2

that have the closest cutoff 99.5. This corresponds to the difference in the share of products purchased with the message "+50" that has 99.5 as the closest cutoff and products with the message "+100". Intuitively, it is hard to imagine a significant effect of treatment at this cutoff, and the difference in means does not reflect such a treatment effect. Instead, this difference reflects that the products below the cutoff are significantly different to the products above the 50 cutoff as shown in the comparison of means in Appendix A.2. As an example, the products below the cutoff are on average significantly more expensive than the products above the cutoff, and a significantly lower share of the products below the cutoff are on campaign.

Figure 4 - Model estimated without bandwidth, graphical illustration

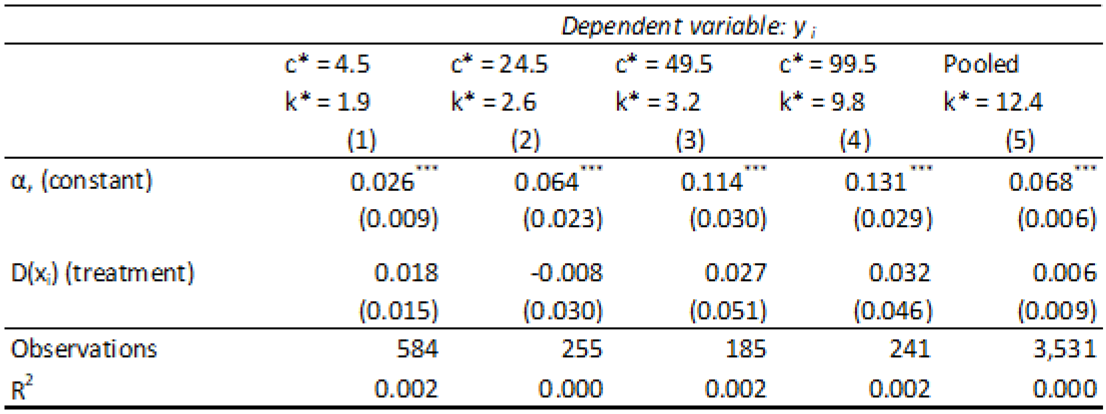


Note: Solid red line is predicted value from estimating baseline model without bandwidths. Dashed red line indicates the predicted 95%-confidence interval (upper and lower bound). The grey dots are the means of the observations (binned distance to cutoff). There is a maximum of 22 bins pr. graph. For "Message cutoff = 99.5" and "Message cutoff = Pooled" the distance to cutoff ranges from -25 to approx. 140,000 (Elgiganten have 140,000 face masks in inventory), due to clear visibility issues we have constrained the visualisation to range from -25 to 25 but the average is taken across the entire inventory range.

Source: Data in Appendix A.1 and own calculations (displayed in appendix A.2)

In Table 2 below, we present the estimation of our baseline model. The estimations (1) through (4) are estimations of the model where the data is restricted to only include observations with the closest cutoff respectively equal to 4.5, 24.5, 49.5, and 99.5. Model (5) is an estimation of model (1) where all the cutoffs are pooled. The optimal bandwidth is calculated as outlined by Imbens and Kalyanaraman (2009). The optimal bandwidth estimated differs across all 5 models due to the included data being different for all 5 models.

Table 2- Estimation of baseline model with an algorithmically decided optimal bandwidth



	Dependent variable: y_i				
	c* = 4.5 k* = 1.9 (1)	c* = 24.5 k* = 2.6 (2)	c* = 49.5 k* = 3.2 (3)	c* = 99.5 k* = 9.8 (4)	Pooled k* = 12.4 (5)
α, (constant)	0.026^{***}	0.064^{***}	0.114^{***}	0.131^{***}	0.068^{***}
	(0.009)	(0.023)	(0.030)	(0.029)	(0.006)
$D(x_i)$ (treatment)	0.018	-0.008	0.027	0.032	0.006
	(0.015)	(0.030)	(0.051)	(0.046)	(0.009)
Observations	584	255	185	241	3,531
R^2	0.002	0.000	0.002	0.002	0.000

Note: *p<0.1, **p<0.05, ***p<0.01, c* indicates the products' closest cutoff that are included in the model, k* is estimated optimal bandwidth.
Source: Data in Appendix A.1 and own calculations (displayed in Appendix A.2)

From the table, we see that there are no statistical significant effects of the treatment on the share of unique products purchased across the 5 estimations presented. Thus, the results suggest a rejection of the hypothesis outlined in the introduction.

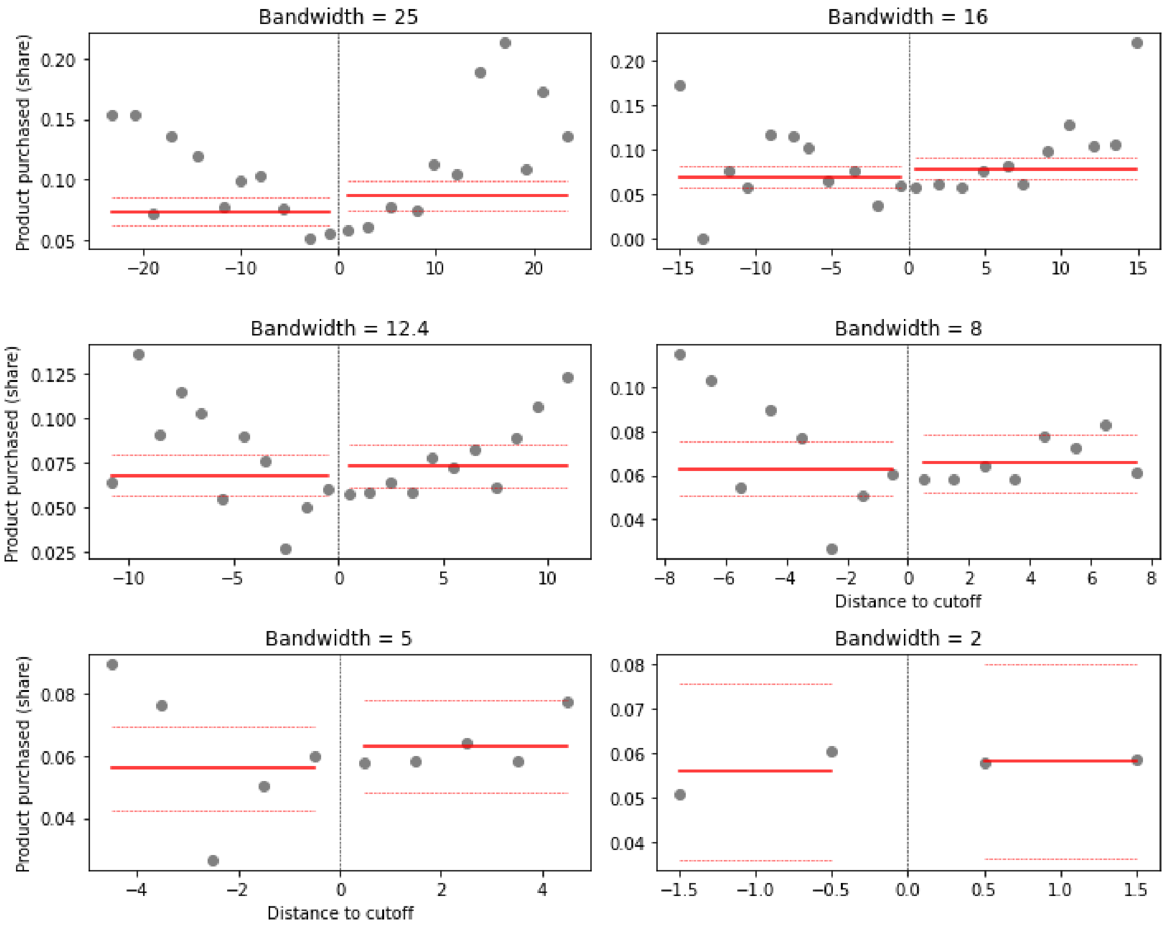
51

Since the average treatment effect is estimated locally around the cutoffs (LATE), conclusions drawn on the insignificant treatment effect can only be based on the particular subset of products included in the estimation. These products might have different properties to products that are far away from the cutoffs. Hence, pushing different inventory messages to customers interested in the products far away from the cutoff (e.g. products with an inventory level above 150) might have a significant impact on customers' purchasing behaviour. This is a limitation of relying on local randomisation for identification of causal effects.

45

Note from estimation (1) through (4) in Table 2 that the optimal bandwidth is increasing with the value of the closest cutoff. This is due to the change in the list of possible values that the running variable can take around the cutoff. As an example, there are fewer possible values of the running variable for products that have 4.5 as the closest cutoff compared to the possible values for products that have 99.5 as the closest cutoff. To see how the bandwidth influences results, we have estimated the model with different bandwidths. In Figure 5, we present a graphical representation of the estimations of our baseline model using the pooled discontinuities with different bandwidths.

Figure 5- Model estimated with pooled discontinuities and different bandwidths



Note: Solid red line is predicted value from estimating the baseline model. the dashed red lines indicate the predicted 95%-confidence interval (upper and lower bound). The grey dots are the means of the observations (binned distance to cutoff). There is a maximum of 22 bins pr. graph. Bandwidth 12.4 is the "optimal" bandwidth, hence an illustration of the estimation in Table 2 (5). **Source**: Data in Appendix A.1 and own calculations

50 In Figure 5, it is clear that the difference in means, the LATE, decreases as bandwidth size decreases. It is possible to modify the baseline model to justify having a high bandwidth by controlling for linear trends on each side of the cutoff. For instance, the relationship between the outcome and the running variable in the figure with bandwidth=25 in Figure 5 appears somewhat linear. We have estimated the model with linear trends, and it does not alter the main takeaways from the baseline estimation in Table 2 (as the figure also suggests for the pooled cutoffs i.e. no "jump" around cutoff).[3]

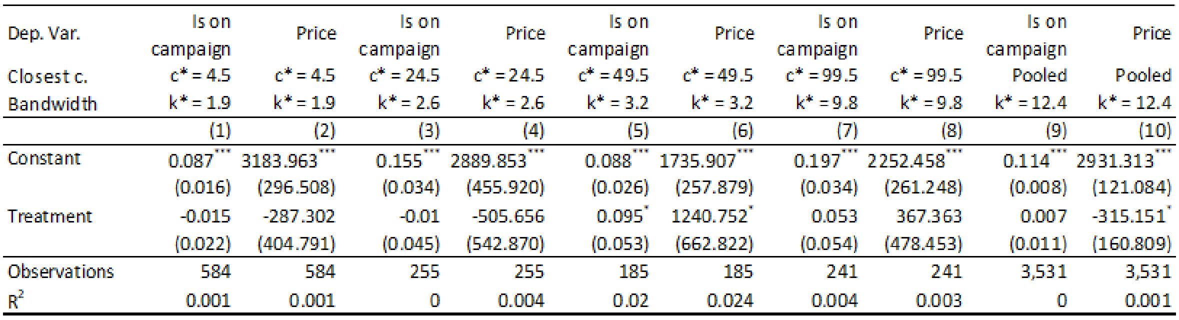
Covariate balance

51 To check the conditional independence of potential outcomes i.e. whether it is a fair assumption that treatment is assigned randomly conditional on the optimal bandwidth, we check for

[3] Model specification and estimation with linear trends is presented in the last part of Appendix A.1.

covariate balance. We do this by regressing the treatment on two predetermined variables namely price and whether the product is on campaign. We use the optimal bandwidths presented in Table 2 when doing the covariate balance check. Results of the covariate balance check are shown in Table 3 below:

Table 3 - Covariate balance check



Dep. Var.	Is on campaign	Price	Is on campaign	Price	Is on campaign	Price	Is on campaign	Price	Is on campaign	Price
Closest c.	c* = 4.5	c* = 4.5	c* = 24.5	c* = 24.5	c* = 49.5	c* = 49.5	c* = 99.5	c* = 99.5	Pooled	Pooled
Bandwidth	k* = 1.9	k* = 1.9	k* = 2.6	k* = 2.6	k* = 3.2	k* = 3.2	k* = 9.8	k* = 9.8	k* = 12.4	k* = 12.4
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Constant	0.087***	3183.963***	0.155***	2889.853***	0.088***	1735.907***	0.197***	2252.458***	0.114***	2931.313***
	(0.016)	(296.508)	(0.034)	(455.920)	(0.026)	(257.879)	(0.034)	(261.248)	(0.008)	(121.084)
Treatment	-0.015	-287.302	-0.01	-505.656	0.095*	1240.752*	0.053	367.363	0.007	-315.151*
	(0.022)	(404.791)	(0.045)	(542.870)	(0.053)	(662.822)	(0.054)	(478.453)	(0.011)	(160.809)
Observations	584	584	255	255	185	185	241	241	3,531	3,531
R^2	0.001	0.001	0	0.004	0.02	0.024	0.004	0.003	0	0.001

Note: *p<0.1, **p<0.05, ***p<0.01 Estimation of $cov_i = \alpha + D(x_i) + \epsilon_i$ imposing the same restrictions on data as the estimation of the baseline model presented in Table 2. c* indicates the products' closest cutoff that are included in the model, k* is the bandwidth.
Source: Data in Appendix A.1 and own calculations (displayed in Appendix A.2)

We find that there are no statistically significant correlations between the treatment and respectively the price and whether the product is on campaign at a 5 pct. significance level. However, there are slight indications of statistical significant correlations between the treatment and covariates, when the data is restricted to products close to the 49.5 cutoff. Around this cutoff both of the covariates are significantly correlated with treatment at a 10 pct. significance level. If there is a relationship between the outcome and the covariates, our assumption of local randomisation is jeopardised: Products on either side of the cutoff would be substantially different and our research design would no longer validly estimate causal effects of the treatment. Our covariate balance check indicates minor possible issues with local randomisation, which we take into account when evaluating the results.

Discussion

In the previous section, we presented the results of our analysis testing the hypothesis that a customer receiving a message that a product has lesser inventory would increase the customers likelihood to buy the product. In our analytical approach, we have not been able to test all facets of this hypothesis due to restrictions of our dataset and restrictions that are implicit in the RDD.

As outlined in the data section, we only have aggregated data on the purchases of a given product on a 24 hour basis. Hence, we measure the inventory 24 hours before the purchase and use this inventory status to create a deterministic function mapping which kind of treatment the customers have received within the 24 hours. However, since the purchase of a product influences the inventory level it might also influence the inventory status message that future customers of a product receive. Due to this limitation of our data we have chosen to model the impact of whether a product was bought or not during the 24 hours, which limits the analysis to 50 only consider the effect of less popular products. To illustrate this, consider a hypothetical product that on average is purchased 3 times every day. This product is probably bought at least one time no matter the message received by the customers. If we hypothesise that a treatment effect exists, the product might be purchased 4 times if there are few in stock or 2 times if there are more. This hypothesised treatment effect will not be captured by our baseline model due to the construction of the outcome variable. Hence, the effect of the message on purchasing behaviour for more popular products is not modelled. Instead of modelling the outcome variable as a dummy, we could construct a variable that consisted of the amount of purchases of the product with different treatments in two ways: Either by doing high frequency scraping of the Elgiganten website or by constructing the amount of purchases of a product that have received different treatments by exploiting the deterministic nature of the treatment.[4] However, since Elgiganten carries large stocks of the products that on average are purchased multiple times daily, the observations are far from the highest cutoff ("+100" products in stock), why we argue that this modification would have a small impact on results.

The same correlation between product quantities and purchases of products is what complicates the interpretation of the (insignificant) treatment effect we estimate. As mentioned in the results section, the LATE is a local effect and this implies that it can only be interpreted as 51 a treatment effect for inventory messages for products that are similar to products close to cutoff. These products constitute a difficult-to-interpret subsample of all of Elgiganten's products.

Our RDD study finds indicative evidence of the lack of a causal impact of changing inventory status messages on the probability that a given product will be purchased a 24 hour period. I.e. we did not find evidence indicating that there is an effect of the digital nudge implemented on the Elgiganten webpage on customers' purchasing behaviour. This is not unexpected as the message about inventory status (the nudge) is small, and if products are not in stock Elgiganten provides a resupplying date, which is often within a short period of time.

We conclude that there is indicative evidence of a lack of an effect while keeping previously mentioned considerations of study and data limitations in mind. These limitations imply that

[4] Since the treatment is a deterministic function of the inventory, it is possible to determine the amount of purchases of a given product that have had different treatments. For instance a product that has 8 in inventory and has been purchased 5 times could be divided into two observations: 3 purchases, where the customer was exposed to the inventory status message "+5" in stock and 2 purchases at "<5". However, the interpretation of an RDD using this outcome variable would not be straightforward, since the timeframe where the inventory status message is displayed differs.

our finding is only a local finding around cutoffs, and that there is a possibility that different decisions around research design and data collection could have yielded other results.

Part 2: Measuring ties among students at DTU through a comparison of two digital networks

Introduction

For this part of the exam, we explore Facebook data and phone call records in order to try to define modes of social relatedness in a university cohort. Scholars have previously studied relations among university students using different types of network data (Kossinets 2009; Ralund and Carlsen 2021;Sekara et al. 2015). Previous data for network analyses was mainly collected through surveys or observation (Wasserman and Faust [1994]2009), but the digital age has brought along multiple new sources of relational data well-suited for comprehensive network analyses. For the present study we delve into the possibilities of digital data and explore how it can be used to understand social relations. We attempt to understand differences in modes of social relatedness among students at The Technical University of Denmark (DTU) by examining their ties in respectively a network of phone calls and Facebook friends. Our research question is:

> *How do two networks of digital data from respectively phone calls and Facebook differ, and how can these differences contribute to an understanding of modes of social relatedness in a cohort of students?*

Our approach for answering this question consists of two steps: The first step is a comparison of the two different networks, respectively a network of students' Facebook-friends and a network of their interactions through phone calls. These turn out to be different with regards to several network statistics including density and degree centrality measures. We argue that the social ties among students in the Facebook-network are predominantly weak, whereas the ties in the network of phone-calls are strong. The second step is an exploration of the distribution of ties in both networks, where we combine the networks to identify four distinct modes of social relatedness among the students, namely: *A. The peripherally connected, B. The closely connected, C. The weakly connected,* and *D. The socially crosscutting.* We then ascribe each student a mode of social relatedness based on their set of ties in each of the two networks. The modes of social relatedness can be considered analytical categories that indicate how students relate to one another, and they constitute one approach of understanding social relatedness through an analysis of digital network data.

Data

51 We will work with publically released data from the Copenhagen Network Study (Sapiezynski et al. 2019).[5] For this study, a range of different digital data types were collected from the 2013 student cohort at The Technical University of Denmark (DTU) (ibid.). The relational data collected corresponds to a one mode network as it only contains one set of actors, the students (Wasserman and Faust 2009:29-31). Students who opted to be study participants were given a smartphone. The researchers then obtained data from phone calls, texts, bluetooth copressences and Facebook friendships between the students in a period of four weeks (ibid.). In our project, we will employ the data on phone calls and Facebook friendships in order to compare and combine the two distinct sets of network relations.

Modifications of the datasets

45 The data was collected for the same set of students across all data types, however nodes only appear in the datasets available to us if they have formed an edge. We modify both datasets in order to work with a consistent set of actors. A majority of students (N=505) appear in both our chosen datasets. A significant proportion of students (N=295) do not appear in the call data. All students were provided with a smartphone, so these are students who did not make any phone calls to other study participants. Thus, we add these students to the network of call data as nodes without any edges. A smaller subset of students (N=31) are only in the call data. Sapiezynski et al. explain that some students did not opt in to the data collection from Facebook (Sapiezynski et al. 2019:4). However, it is also possible that some of the students had no Facebook friends among the study participants or no Facebook account. We are not able to distinguish between students not opting in and students without Facebook friends in the cohort, therefore we choose to remove these students from the dataset in order to preserve a consistent set of actors. After these modifications the two datasets include the same 800 students.

The call data contains directional edges between each caller and callee. However, we choose to transform these edges to be non-directional as we operationalise the phone calls to represent social ties which we assume to be mutual and thus undirected.

Step 1. Comparison of the two networks

50 In the first step of our analysis, we estimate network statistics operationalised as different measures of sociality for each of the two networks. We measure social cohesion and social stratification of the cohort overall, and we estimate each student's individual social connection

[5] The public data from The Copenhagen Network Study can be retrieved from the website Netzschleuder through this link: https://networks.skewed.de/net/copenhagen

to the rest of their cohort. We then compare these estimations in order to understand how the two networks differ with regards to types of relations between the students.

Social cohesion - Network density

51 In this project, we view social cohesion as the degree to which students have formed relations and thus are connected to each other. We can imagine a hypothetical spectrum of social cohesion: On one end of this spectrum would be an atomistic structure, where no students in the cohort established ties with their peers, nodes without edges. In the other end there would be a completely connected network, where each student had ties to every other student in their cohort. We interpret the density of each network as the levels of social cohesion within the student cohort. Density is a standardised way of measuring how connected the nodes in a network are, and it is a recommended and often used measurement for the “knittedness” of a network (Blau 1977; Wassermann and Faust 2009:181). For a non-weighted and non-directed network, this metric is computed using the following formula:

$$density \;=\; \frac{L}{g(g-1)/2}$$

45 L is the total number of edges and g is the total amount of nodes. $g(g\,-\,1)/2$is the total amount of possible edges in the network. Density is thus the share of present edges compared to the number of all possible edges in a network. The metric is standardised as it is a fraction that ranges from zero to one for all networks. However, in social networks density is sensitive to the size of the network as it becomes increasingly difficult for each actor to establish and maintain ties with all other actors in the network (Scott 1988:114-115). The two networks have the same size, hence, a direct comparison of density is possible in this case.

Each person’s social connection to cohort - Nodal degree

50 We want to measure how socially connected each student is to their cohort on an individual level in order to construct a typology of different modes of social relatedness. We view the number of connections a student has to other students as an operationalisation of how socially connected that student is to their cohort. The number of social relations each student has formed might differ, and this can be reflected in the nodal degree. The nodal degree $d(n_i)$ is the total number of edges that are incident with a given node (Wasserman and Faust 2009:163,178). By using the symmetrical adjacency matrix, A, where both rows and columns represent the same 800 nodes in the datasets, we can compute the nodal degree for each node in each network as follows (ibid.):

$$d(n_i) = \sum_j \mathrm{A}_{\mathrm{i,j}} = \sum_i \mathrm{A}_{\mathrm{j,i}}$$

51 where *i* is a node and *j* is a node. As reflected in the equation, the sum of the rows for a given column is the same as the sum of the columns of a given row. This follows from the properties of the adjacency matrix: Each row is equal to one if the node is connected with an edge to that node and zero if not. Similarly, each column value is equal to one if the row is connected to the node in the column. Therefore, we end up with the nodal degree for a given node by summing either all the rows for a given column or all the columns for a given row. Since the two networks contain the same people there is no need to standardise this measure to compare the nodal degrees between the two networks.

Social stratification - Group degree centralisation

45 For this project, we view social stratification as the inequality in the distribution of social ties among the students. Is there a strong group of students who hold nearly all of the social relations while the rest of the students have relatively few social ties, or do students have a more equally distributed number social relations in the cohort? As a measure of social stratification, we use the nodal degree to calculate the group degree centralisation. The standardised group degree centralisation,C_D, and it is a measurement of the sum of differences between the largest observed nodal degree in the network and each observed nodal degree value in relation to the number of possible edges (Wasserman and Faust 2009:180). Thus, it is a credible measure to use for interpreting the stratification of social ties. Formally, the group degree centralisation measure is:

$$C_D = \frac{\sum_{i=1}^{g}[C_D(n^*) - C_D(n_i)]}{[(g-1)(g-2)]}$$

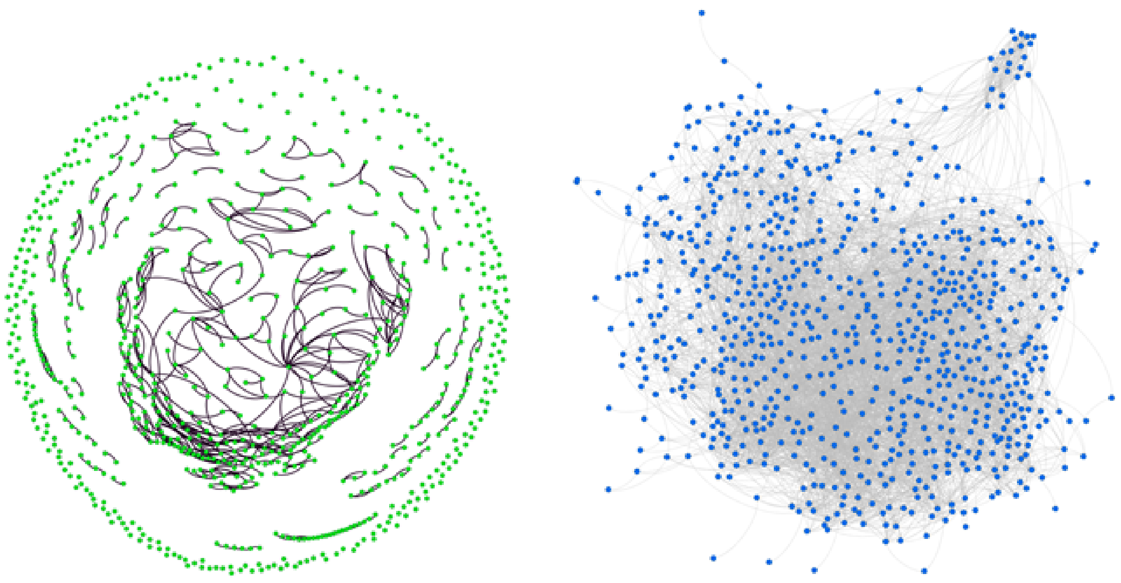
$C_D(n_i)$corresponds to the nodal degree $d(n_i)$ for the node n_i and $C_D(n^*)$is the largest observed nodal degree. The measure is standardised by dividing with the total amount of possible edges (g-1)(g-2), and it returns a value in an index between 0-1 allowing for easy comparison.

Results

50 First, we look at the social cohesion of the network of students by estimating the network density. The density of the Facebook network is 0.0201 whereas it is 0.0017 for the call network as calculated in Appendix B.2. As mentioned, we would expect relatively low density scores for a social network of 800 people, as each person is limited in how many ties they are willing and able to form (Scott 1988:114-115). The density score is about twelve times as large in the

Facebook network compared to the call network. Since the two networks contain the same people, we argue that this shows that the threshold for establishing a Facebook friendship is lower than for making a phone call. This falls in line with earlier research from Manago et al. who find that Facebook facilitates the expansion of especially distant aqcuantainships (Manago et al. 2012). It is also in line with an analysis by Ralunds and Carlsen of ties among Danish university students where they state that "(...)facebook friendships are often formed very casually, with even one time meetings resulting in a friend request (and acceptance). " (Ralund
51 and Carlsen 2021:12). Visualisations of two distinct networks using ForceAtlas2 are shown below in Figure 6, and here the differences in density is clear. Here the Facebook network is illustrated as a "hairball"-structure where all nodes are connected, whereas the call network has multiple layers showing the different actor degree centralities. Call relations are less common in the network as they are an active act of communication, which we interpret as indicating a closer relationship between actors. Thus, we treat Facebook friendships as signifying that an acquaintanceship has formed between two students, whereas, call relations are treated as closer ties corresponding to friendships

Figure 6: Networks of ties in call data and Facebook data



Note: Left - Network of phone calls. Right - Network of Facebook relations. **Source:** Data from Appendix B.1 and own calculations in Appendix B.2.

45 We calculate the nodal degree for each actor and use it to find the group degree centralisation on network level in order to measure the stratification of the ties in the two networks. The group degree centralisation for the Facebook network is 0.1065 while it is 0.0183 for the call network. This shows that Facebook friendships are more stratified than phone calls, we interpret this as

acquaintanceships (Facebook friendships) being more unequally distributed than closer relations (phone calls) among the cohort. This entails that a smaller subset of students have developed a larger share of weak ties as compared to strong ties.

Step 2: Defining modes of social relatedness

50 As we have now shown some of the main characteristics of each of the two networks, we proceed to combine the networks in order to define modes of social relatedness by drawing on these characteristics. This is illustrated in Table 4 which shows four groups that we interpret as indicators of distinct modes of social relatedness among the students.

Table 4 - Modes of social relatedness

	Few strong ties (Few phone calls)	**Many strong ties** (Many phone calls)
Few weak ties (Few Facebook friends)	A. The peripherally connected	B. The closely connected
Many weak ties (Many Facebook Friends)	C. The weakly connected	D. The socially crosscutting

Note: The table displays the four modes of social relatedness based on number of Facebook ties and phone calls

This division of students into four distinct groups is a construct that works as a way to differentiate between approaches of engaging in social relations reflected in the networks. There are multiple limitations with regards to these rough generalisations which will be discussed in a later section. We now describe each of the different modes depicted in the table.

A. The peripherally connected

51 We interpret Group A to represent a mode of peripheral relatedness as students in this group have only few ties in both the network of weak ties (Facebook network) and the network of strong ties (phone calls). Thus, this mode is characterised by few friends and acquaintances among peers which might indicate low participation in social activities.

B. The closely connected

The students in Group B have many strong ties but few weak ties. This indicates that the students have friends among their peers but they are less inclined in participating in more loose acquaintanceships. We thus interpret their main mode of relatedness to be characterised by predominantly participating in closely connected friendship groups.

C. The weakly connected

The students in Group C have many weak ties and few strong ties. We have assigned the mode of social relatedness characterizing this group to be weakly connected as they tend to have a lot of loose acquaintances among their peers rather than engaging in close friendships.

D. The socially crosscutting

We interpret Group D as being characterised by a mode of being socially crosscutting as they
45 have both many strong and many weak connections. This indicates that students in this group
are highly socially engaged in their study environment as they manage to have both friends and
acquaintances among their peers.

Measures of modes of social relatedness

In order to assign each node to a mode of social relatedness, we find the *nodal degree* for each node and compare it to *mean nodal degree*. The mean nodal degree is a statistic that summarises the degree of all nodes by denoting the average nodal degree in a network (ibid.). The mean nodal degree is denoted as:

$$\bar{d} = \frac{\sum_{i=1}^{g} d(n_i)}{g} = \frac{2L}{g}$$

In the above equation, g is the total number of nodes in a network and L is the total number of
50 edges (ibid.:100-101). Each edge is connected to two nodes, and thus $2L$corresponds to the sum
of nodal degrees in a network.

We find the mean nodal degree for each network, and use it to compare with the nodal degree for each node in both networks. We have depicted our criteria for assigning a node to a specific group in Table 5 below.

	Below mean nodal degree in network of **phone calls**	Above mean nodal degree in network of **phone calls**
Below mean nodal degree in network of **Facebook ties**	Group A (N=357)	Group B (N=159)
Above mean nodal degree in network of **Facebook ties**	Group C (N=147)	Group D (N=137)

Note:: The matrix displays how nodes are divided into groups representing the four modes of social relatedness

If a node has a nodal degree below the mean in both networks, the node is assigned to Group A. If a node has degree below mean in the network of Facebook ties and above mean in the call network, the node is assigned to Group B. If a node with a nodal degree above the mean in the Facebook network and below the mean for the call network is assigned to Group C, and if a node with a nodal degree above the mean in both networks is assigned to Group D.

A weighted multigraph

We can visualise a combination of the two networks in a *multigraph* using colors to indicate the
51 different groups. A multigraph is a graph that allows for more than one edge between two nodes
and thus more than one type of relation between a set of individuals (Wasserman and Faust

2009:145-146). Hence, we can construct a network consisting of both Facebook friend ties and phone call ties between the students.

We have argued throughout Step 1 and 2 that ties in the Facebook network are weaker than ties in the call network. We operationalise this finding by constructing a weighted network. A weighted network is a network where the edges have different non-negative integer values (Wasserman and Faust 2009:141). Thus, we assign larger weights to the ties in the call network than to the ties in the Facebook network to illustrate that these ties are stronger. A visualisation of the weighted multigraph allows us to further explore and identify patterns for the defined modes of social relatedness among the students.

Results

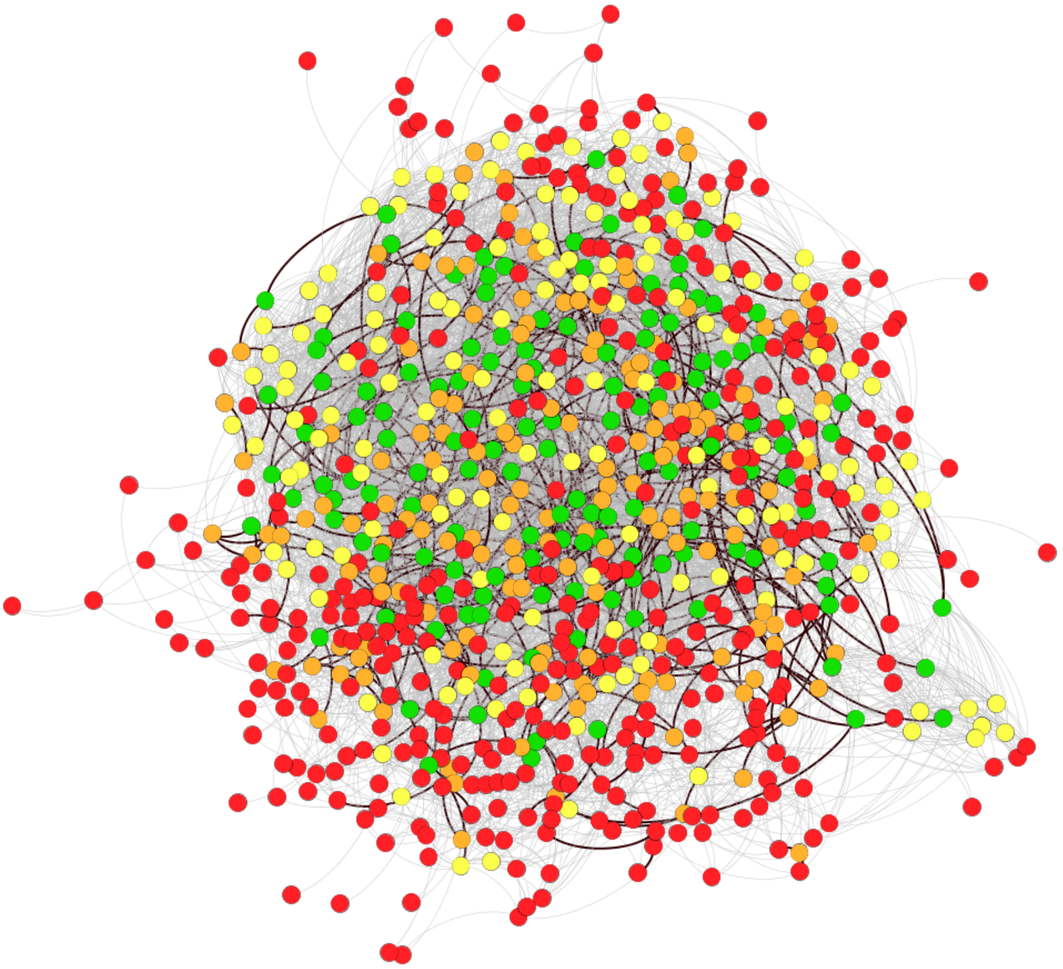
For each node we calculate the nodal degree, and we find the highest values to be 101 in the Facebook network and 16 in the phone call network. Furthermore, the mean nodal degree for the Facebook network is 16.05 and 1.38 for the call network. Each student has on average approximately 16 weak Facebook ties and between 1-2 strong call ties within the cohort.

After calculating the mean nodal degrees, we proceed to find the modes of social relatedness. 357 (44,6%) students are categorised as belonging to Group A, 159 (19,8%) belonging to Group B, 147 (18,3%) belonging to Group C, and 137 (17,1%) belonging to Group D. This means that just below half of the students have few friends and acquaintances among their peers.

In Figure 7, the multigraph is illustrated using ForceAtlas2 (Jacomy et al. 2014). Call edges are illustrated with thick black lines, whereas Facebook friendships are grey thin lines. The color of the nodes indicates the group to which the student is assigned.

Figure 7: The weighted multigraph

● Group A - the peripherally connected
● Group B - the closely connected
● Group C - the weakly connected
● Group D - the socially crosscutting



Note: Multigraph containing a combined network of the Facebook network and the call network. The coloured labels represent the different modes of social relatedness.

In Figure 7, stratifications of the groups can be identified: The peripherally connected seem to
50 cluster in the periphery, whereas the socially cross cutting are clustered in the center. In the
section below, we argue that the clustering in the periphery might indicate a vulnerability to be
detached from the social life in the study cohort.

Consequences of modes of relatedness

51 In order to interpret possible consequences of the four modes of social relatedness with regards to students' retention rates and social satisfaction, we draw on a paper by Ralund and Carlsen about benefits of different network structures at 44 danish university programs (Ralund and Carlsen 2021). They describe *individualised network structures* as consisting of far-flung, demanding, and sparsely knit relations which tend to be unstable and lead to a higher probability of low retention rates for individuals in the network (ibid.:2,27). *Grouped networks* on the other hand are characterised by closely knit ties between individuals in close groups which lead to higher social satisfaction and high retention rates (ibid.:10,27). Drawing on these findings, we argue that the strong ties of the call network can be interpreted as indicators of more closely knit ties that are likely to positively influence whether students are thriving both socially and with regards to their education. We thus expect the peripherally connected (Group A) and the weakly connected (Group C) to reflect modes of relatedness where individuals experience lower retention and lower social satisfaction than the modes of relatedness characterised by higher rates of strong ties (Group B and Group D). Further analysis including more data on the students could allow us to test this theory.

Discussion

45 operationalisation is founded on assumptions that have some limitations. In order to interpret Facebook ties as weak relations and phone call ties as strong relations, we have assumed that the different ways of using these media reflect certain modes of social relatedness. Nevertheless, there are most likely a variety of practises employed when it comes to media and social relations, and as an example it is likely the case that some students are close friends without having called each other on the phone. As such, call data may be more accurate when estimating the presence of a social relation compared to estimating the absence of one.

One way to refine our operationalisation that phone calls indicate friendship ties could be to weigh the call data by exploiting the fact that it is directional data. This could be done by restricting assignment of strong ties to instances where a set of nodes are connected by multiple call ties in each direction. However, in this case we do not measure the strength of a friendship, but merely the presence of two different types of ties.

Another limitation of our operationalisation of the data is that some students are not connected to anyone in the Facebook network which might be due to lack of acquaintances, but it could also be because of Facebook inactivity or declined consent to the data collection, and these individuals might thus still have weak ties among their peers. As we do not know the reason for the absent ties in the Facebook network, we chose to discard these 31 individuals while being aware that this might create a bias in our analyses leading to estimated values of density and mean degree for the Facebook network that are higher than they would otherwise have been.

50 Though digital data brings new opportunities for social data science, the process of operationalisation is often challenged (Salganik 2017:2.3.4). This is because the data is often incomplete as it was initially constructed for other purposes (ibid.). However, we argue that the identified differences in the structure of the Facebook data and the phone records reveal interesting information about how the two media are used by students. In line with previous studies, we further argue that our interpretation of the media connecting respectively weak and strong ties are not too far fetched, and might reveal overall tendencies about how students relate to one another.

An extension to this study could be to identify the k-cores of the multigraph in order to explore to which extent the identified groups of modes of social relatedness cluster together and whether this is an expression of homophily within the cohort of students.

Conclusion

51 We have answered our research question in two steps: First we have examined the differences of the two networks of digital data, where we found that call ties represent closer relations than Facebook ties, and that access to loose Facebook ties are more stratified than close call ties. Secondly, we used this finding to explore modes of social relatedness, which we illustrated in a combined multigraph, that shows how a large part of the students are posited in the periphery of the network with only a few close ties. Nevertheless, there are limitations in the study that must be taken into consideration when the results are interpreted, especially when considering the call data. It is important for us to stress that the constructed modes of social relatedness are merely considered indicators of a way of balancing distribution of social ties. We do not consider the modes of social relatedness to represent reality on a one-to-one basis but rather as analytical categories that can be a prism for looking at social relatedness.

Literature

Angrist, Joshua D. and Jörn Steffensen Pischke. 2009 *Mostly Harmless Econometrics*. Princeton: Princeton University Press.

Angrist, Joshua and Jörn-Steffen Pischke. 2015. *Mastering Metrics: The Path from Cause to Effect.* Princeton: Princeton University Press

Blau, Peter Michael. 1977. *Inequality and heterogeneity: A primitive theory of social structure*. New York: Free Press.

Jacomy Mathieu, Tommasso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software". In *PLoS ONE.* Volume: 9, Issue: 6

Kossinets, Gueorgi and Duncan Watts. 2009. "Origins of Homophily in an Evolving Social Network". In *American Journal of Sociology.* Volume:115, Issue:2.

Imbens, Guido and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator" In *The Review of Economic Studies*. Volume: 79, Issue: 3.

Manago, Adriana M., Tamara Taylor, and Patricia M. Greenfield. 2012. "Me and My 400 Friends: The Anatomy of College Students' Facebook Networks, Their Communication Patterns, and Well-Being." In *Developmental Psychology.* Volume: 48, Issue: 2.

Ralund, Snorre and Hjalmar Carlsen. 2021. "Cross cutting or grouped networks - What network structures support social integration and wellbeing in adolescent collectives?" *Working Paper*

Salganik, Matthew. 2017. *Bit by Bit: Social Research in the Digital Age.* New Jersey: Princeton University Press. Available on: https://www.bitbybitbook.com/en/1st-ed/introduction/

Sapiezynski, Piotr, Arkadiusz Stopczynski, David Dreyer Lassen, and Sune Lehmann. 2019. "Interaction Data from the Copenhagen Networks Study". In *Scientific Data.* Volume: 6, Issue:1

Scott, John. 1988. "Social Network Analysis". In *Sociology.* Volume: 22, Issue 1.

Sekara, Vedran, Arkadiusz Stopczynski, and Sune Lehmann. 2016. "Fundamental structures of dynamic social networks." In *Proceedings of the national academy of sciences.* Volume: 113, Issue:36.

Wasserman, Stanley, and Katherine Faust. 2009[1994]. *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.