Mobile Phone Datasets for Social Good

Challenges and Opportunities

Leo Ferres
UDD & Telefónica R+D, Santiago, Chile
ISI Foundation, Torino, Italia
lferres@udd.cl
FB, IG, Tw, GH: @leoferres

ISI Fellowship Ceremony
Torino, Italia, October 18, 2019
https://github.com/leoferres/isi_fellowship_19

s2019-09-26 12:09:37 -0300 - e:

About me

Available online at www.sciencedirect.com

ScienceDirect

Theoretical Computer Science 00 (2017) 1–24

Theoretical Computer Science

ELSEVIER

Parallel Construction of Succinct Trees☆,☆☆

José Fuentes-Sepúlveda[a,*], Leo Ferres[b], Meng He[c], Norbert Zeh[c]

[a]Department of Computer Science, Universidad de Chile, Santiago, Chile
[b]Faculty of Engineering, Universidad del Desarrollo & Telefónica I+D, Santiago, Chile
[c]Faculty of Computer Science, Dalhousie University, Halifax, Canada

Abstract

Succinct representations of trees are an elegant solution to make large trees fit in main memory while still … operations in constant time. However, their construction time remains a bottleneck. We introduce two … improve the state of the art in succinct tree construction. Our results are presented in terms of *work*, the ti… parallel computation using one thread, and *span*, the minimum amount of time needed to execute a parall… amount of threads. Given a tree on n nodes stored as a sequence of balanced parentheses, our first algorith… representation with $O(n)$ work, $O(\lg n)$ span and supports a rich set of operations in $O(\lg n)$ time. Our sec… the query support. It constructs a succinct representation that supports queries in $O(c)$ time, taking $O(n$ + … and $O(c + \lg(\frac{n c}{\ldots}))$ span, for any positive constant c. Both algorithms use $O(n \lg n)$ bits of working spac… up to 64 cores on inputs of different sizes, our first algorithm achieved good parallel speed-up. We also p… takes $O(n)$ work and $O(\lg n)$ span to construct the balanced parenthesis sequence of the input tree requir… construction algorithm.

Keywords: Succinct Data Structure, Succinct Tree Construction, Multicore, Parallel Algorithm

Knowl Inf Syst
DOI 10.1007/s10115-016-1000-6

REGULAR PAPER

Parallel construction of wavelet trees on multicore architectures

José Fuentes-Sepúlveda[1] · Erick Elejalde[1] · Leo Ferres[2] · Diego Seco[1]

Received: 16 September 2015 / Revised: 20 July 2016 / Accepted: 26 September 2016
© Springer-Verlag London 2016

Abstract The wavelet tree has become a very useful data structure to efficiently represent and query large volumes of data in many different domains, from bioinformatics to geographic information systems. One problem with wavelet trees is their construction time. In this paper, we introduce two algorithms that reduce the time complexity of a wavelet tree's construction by taking advantage of nowadays ubiquitous multicore machines. Our first algorithm constructs all the levels of the wavelet in parallel with $O(n)$ time and $O(n \lg \sigma + \sigma \lg n)$ bits of working space, where n is the size of the input sequence and σ is the size of the alphabet. Our second algorithm constructs the wavelet tree in a domain decomposition fashion, using our first algorithm in each segment, reaching $O(\lg n)$ time and $O(n \lg \sigma + p\sigma \lg n / \lg \sigma)$ bits of extra space, where p is the number of available cores. Both algorithms are practical and report good speedup for large real datasets.

Keywords Succinct data structure · Wavelet tree construction · Multicore · Parallel algorithm

Fast and Compact Planar Embeddings☆,☆☆

Leo Ferres[a], José Fuentes-Sepúlveda[b,c], Travis Gagie[d,c,*], Meng He[e], Gonzalo Navarro[b,*]

[a]Faculty of Engineering, Universidad del Desarrollo & Telefónica I+D, Santiago, Chile
[b]Department of Computer Science, University of Chile, Santiago, Chile
[c]Center of Biotechnology and Bioengineering, University of Chile, Santiago, Chile
[d]School of Computer Science and Telecommunications, Diego Portales University, Santiago, Chile
[e]Faculty of Computer Science, Dalhousie University, Halifax, Canada

Abstract

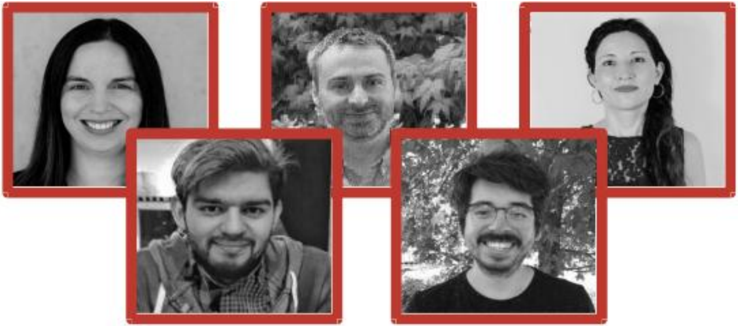
There are many representations of planar graphs but few are as elegant as Turán's (1984): it is simple and practical, uses only four bits per edge, can handle multi-edges and can store any specified embedding. Its main disadvantage has been that "it does not allow efficient searching" (Jacobson, 1989). In this paper we first show how to add a sublinear number of bits to Turán's representation such that it supports fast navigation, and then give a linear-work parallel construction algorithm for the resulting data structure that runs in $O(m/p)$ expected time, where m is the number of edges, p is the number of processors and $p \ll m$. This is the first linear-work and practical parallel algorithm that can encode an embedding of a connected planar graph compactly. We also provide an experimental study of our parallel algorithm and prove that it has good scalability and low memory consumption. Additionally, we describe and test experimentally queries supported by the compact representation.

Keywords: Planar embedding, Compact data structures, Parallel algorithm

☆A previous version of this paper appeared in the 15th Algorithms and Data Structures Symposium (WADS 2017)[8]
☆☆The second and fifth authors received travel funding from EU grant H2020-MSCA-RISE-…

Figure 1: Leo's 2016

About me



Universidad del Desarrollo
Facultad de Ingeniería
Instituto Data Science

Telefónica
Investigación y Desarrollo Chile

Figure 2: The Institute of Data Science, Engineering and TEF R&D

About me


>> # Introduction

- Can we now process 2.1bn records of data?

Introduction

Cut scene to 2018,

- 7.9bn SIM conns, 5.1bn unique mobile subs (2,455,150 x 1.3 in SCL alone), 3.6bn internet users[1]

[1] https://www.gsma.com/r/mobileeconomy/

Introduction

Cut scene to 2018,

- 7.9bn SIM conns, 5.1bn unique mobile subs, 3.6bn internet users
- $\sim$50% of web traffic generated by mobile phones[2],

[2] `https://hostingfacts.com/internet-facts-stats/` (Statista, really)

Introduction

Cut scene to 2018,

- 7.9bn SIM conns, 5.1bn unique mobile subs, 3.6bn internet users
- $\sim$50% of web traffic generated by mobile phones
- computational power for the analysis of VLDBs,

Introduction

Cut scene to 2018,

- 7.9bn SIM conns, 5.1bn unique mobile subs, 3.6bn internet users
- ~50% of web traffic generated by mobile phones
- computational power for the analysis of VLDBs,
- fair-to-good spatial granularity (1,742 towers)[3],

[3] Towers in SCL (1988-2015), see `https://youtu.be/7kZV892QGe4`

Introduction

Cut scene to 2018,

- 7.9bn SIM conns, 5.1bn unique mobile subs, 3.6bn internet users
- $\sim$50% of web traffic generated by mobile phones
- computational power for the analysis of VLDBs,
- fair-to-good spatial granularity,
- (real-time?) temporally fine-grained datasets

Introduction

Cut scene to 2018,

- 7.9bn SIM conns, 5.1bn unique mobile subs, 3.6bn internet users
- $\sim$50% of web traffic generated by mobile phones
- computational power for the analysis of VLDBs,
- fair-to-good spatial granularity,
- (real-time?) temporally fine-grained datasets
- Ecologically valid

Introduction

New quantitative ways of looking at critical social issues:

- gender
- segregation, employment and poverty
- (epidemics) and displacement
- land use
- news consumption

Prelims: Chile and SCL

Figure 3: Chile, 5000km scale

Prelims: Chile and SCL

Figure 4: Chile, 100km scale

Prelims: Chile and SCL

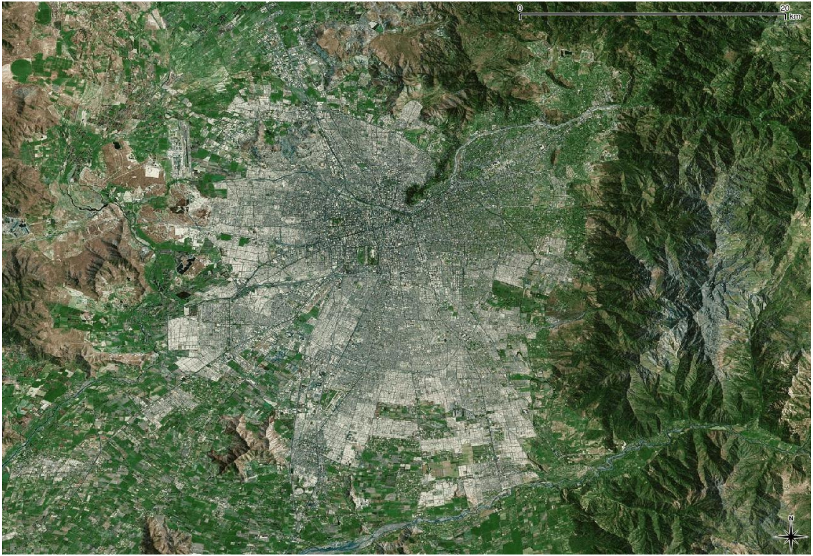


Figure 5: SCL, 20km scale

Prelims: Chile and SCL

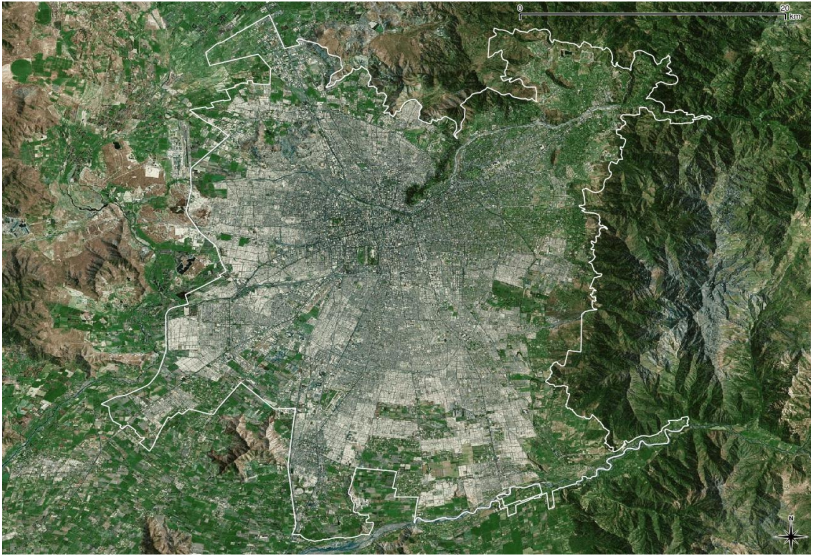


Figure 6: SCL, 20km scale, urban

Prelims: Chile and SCL

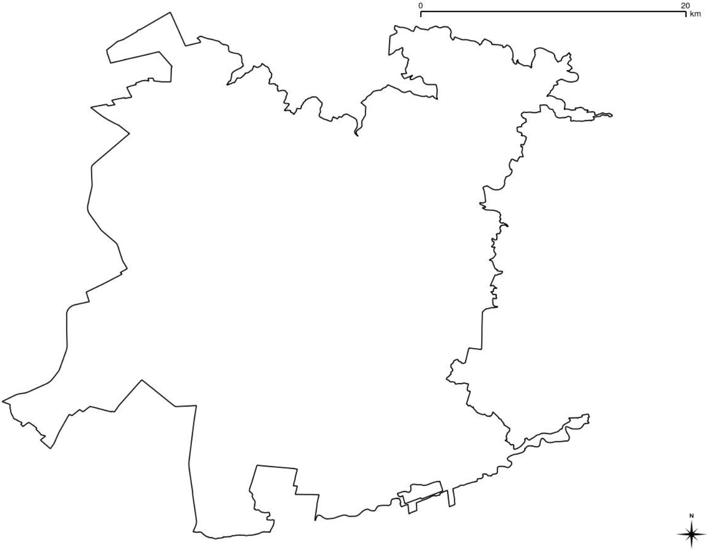


Figure 7: SCL, 20km scale, urban, digital

Prelims: Antennas

Figure 8: A cell tower with the BTS[4]

[4] sciencedirect.com/topics/computer-science/cellular-phone

Prelims: Antennas

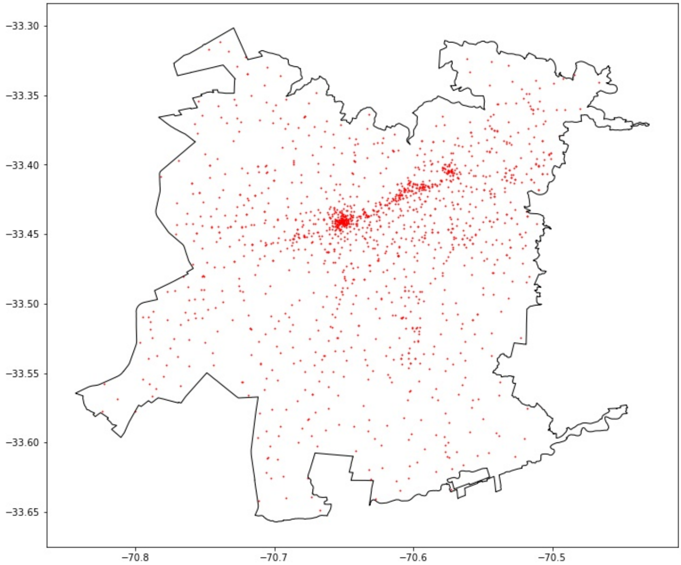


Figure 9: Towers in SCL

Prelims: Antennas

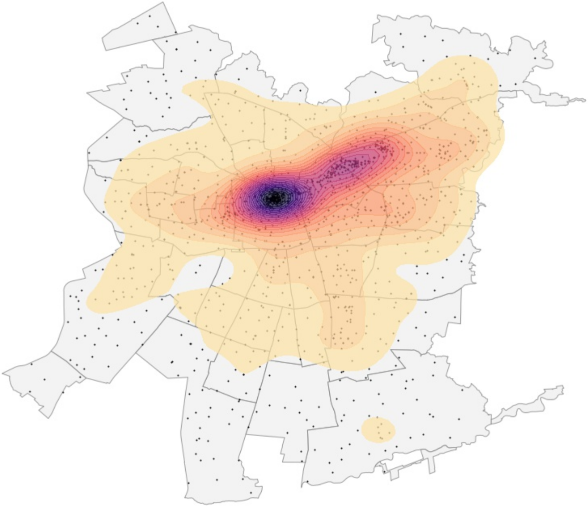


Figure 10: KDE over SCL towers, $n = 20$, @Mao2015

Prelims: HDI

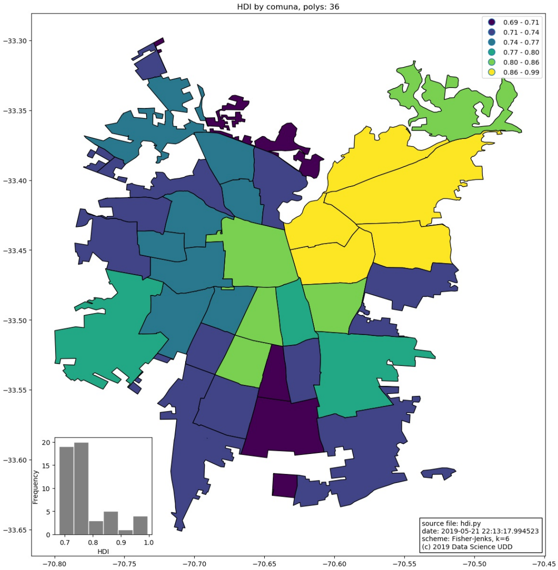


Figure 11: HDI

Prelims: Antennas

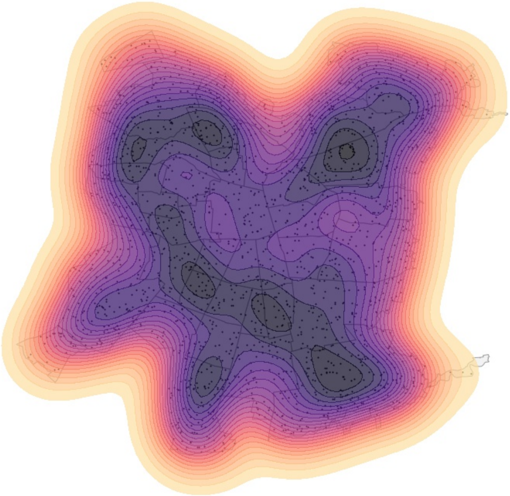


Figure 12: KDE over random SCL towers, $n = 20$

Telephony streams

- **CDR** (Call Detail Record) $\equiv \langle n_a, n_b, t_a, t_b, d, r \rangle$

Telephony streams

- **CDR** (Call Detail Record) $\equiv \langle n_a, n_b, t_a, t_b, d, r \rangle$
- **XDR** (eXtended Detail Record) $\equiv \langle n, t, d, k \rangle$

Telephony streams

- **CDR** (Call Detail Record) $\equiv \langle n_a, n_b, t_a, t_b, d, r \rangle$
- **XDR** (eXtended Detail Record) $\equiv \langle n, t, d, k \rangle$
- **CP** (Control Plane) $\equiv \langle n, t, d, e_1, e_2, \ldots e_n \rangle$

Telephony streams

- **CDR** (Call Detail Record) $\equiv \langle n_a, n_b, t_a, t_b, d, r \rangle$
- **XDR** (eXtended Detail Record) $\equiv \langle n, t, d, k \rangle$
- **CP** (Control Plane) $\equiv \langle n, t, d, e_1, e_2, \ldots e_n \rangle$
- **DPI** (Deep Packet Inspection) $\equiv \langle n_a, t_a, d, k, p \rangle$

CDR: Mobility & Gender

Q[5]: Do we observe similar mobility patterns across gender (in the presence/absence of public transport)?

- **Definition**: An individual i "moves more" than an individual j iff $S^i > S^j$, where S is Shannon's entropy over each individual's i set of all visited places L:
- **Definition**: A "place" $l \in L$ is a 1Km2 cell in a square grid where there is at least one cell tower.

[@Adeel_201 ($<$), @Psylla_2017 ($>$), but @Song_2010 ($=$)]

[5]Study funded by Data2x at the United Nations: ISI Foundation, GovLab, UNICEF, NYU, Digital Globe and us (IDS/UDD/TEF). (Happy to see you here, Stefaan!)

CDR: Mobility & Gender: Datasets

CDR:

- Period:
  - June-August, 2016 (3 months)
  - 2,148,132,995 rows (CDRs, calls), 1.06TB, with GENDER and SEG
  - **372,152** individuals, **50.9%** female

CDR: Mobility & Gender: GTFS public transport

- number of reachable stations
- average velocity to reach other nodes in the network

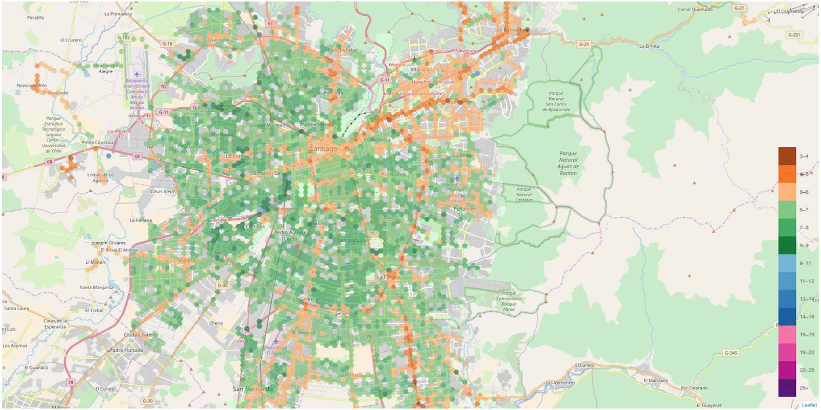


Figure 13: Accessibility map from GTFS public transport data

CDR: Mobility & Gender: Inequality at the city scale

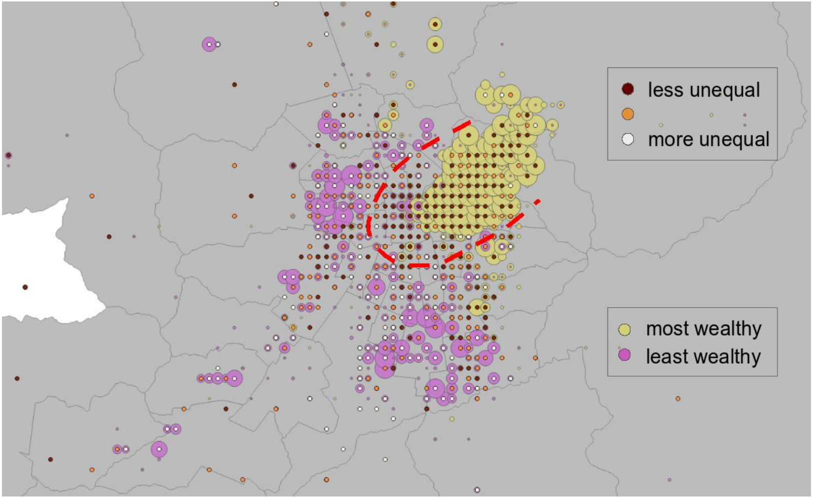
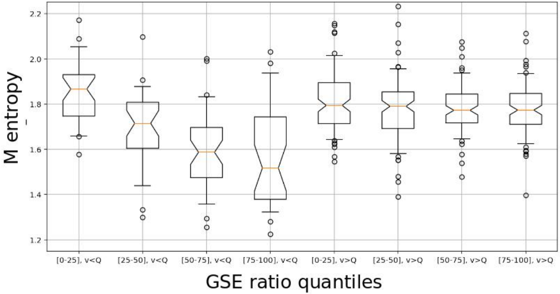
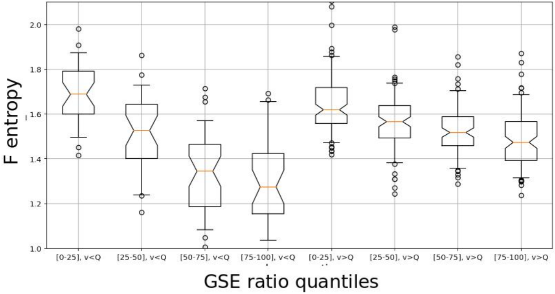


Figure 14: A gender inequality index

CDR: Mobility & Gender: Access to public transport





CDR: Mobility & Gender: Conclusions

- Mobility is (strongly) gendered,
- there are gender differences in mobility patterns when it comes to lack of public transportation,
- we need a smarter way to think about mobility, particularly when it comes to gender

Many more results here: https://arxiv.org/abs/1906.09092

CDR: Mobility & Gender: Conclusions

In any case, TAKE HOME:

- without data equality, there's no gender equality

XDR

Two studies, in order of appearance:

- Graells-Garrido, Ferres, Caro and Bravo. **The effect of Pokémon Go on the pulse of the city: a natural experiment**. EPJ Data Science (2017) 6:23.
  - **Research question:**. What happened to the city after the launch of Pokémon Go?
- Beiró, Bravo, Caro, Cattuto, Ferres and Graells-Garrido. **Shopping mall attraction and social mixing at a city scale**. EPJ Data Science

2018. 7:28.

  - **Research question:**. Given their prominence in the city, and the segregation of Santiago, do Shopping Malls function as social mixers?

XDR: PoGo: datasets

- Period:
  - Jul 27-Aug 2, 2016; Aug 4-7, 2016
  - **142,988** devices active all days, plus
- Origin-Destination Survey
- Ingress Pokestops (Pokemon POIs)

XDR: PoGo: Connections

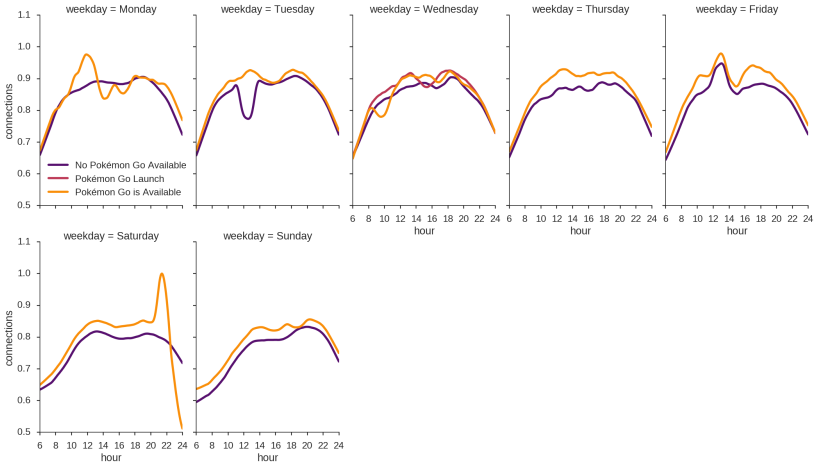


Figure 15: The pulse of the city (floating population profiles) one week before and after the launch of Pokémon Go in Santiago (3rd August).

XDR: PoGo: Results (Incidence Rate Ratio)

Time window	Max IRR	Time of Max IRR
6:34-6:47	1.062	6:40
7:07-7:18	1.056	7:11
7:37-7:46	1.054	7:42
7:48-7:48	1.047	7:48
9:35-9:43	1.060	9:40
10:27-10:41	1.077	10:34
10:53-11:18	1.071	11:07
11:58-12:46	1.138	12:31
13:06-13:09	1.051	13:08
15:36-15:51	1.058	15:50
16:17-16:21	1.052	16:19
18:30-18:34	1.052	18:31
19:42-19:45	1.051	19:43
21:24-22:12	1.096	21:38
22:22-22:25	0.955	22:25
22:44-22:52	0.954	22:52
23:09-23:21	0.954	23:09
23:57-23:59	1.057	23:59

Figure 16: Time windows the PoGO effect was significant (11:58-12:46, and 21:24-22:12)

XDR: PoGo: Results (geoloc'ed)

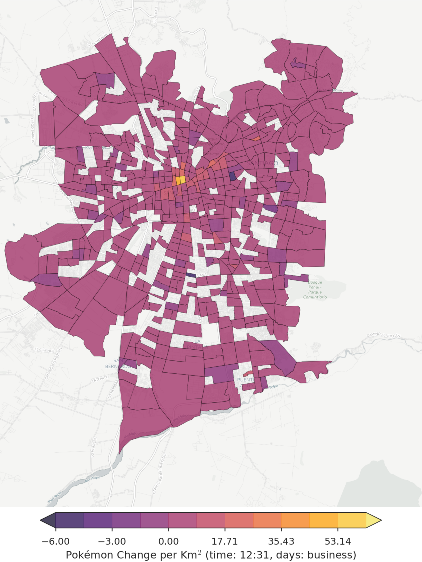


Figure 17: 12.31 Business day

XDR: PoGo: Results (geoloc'ed)

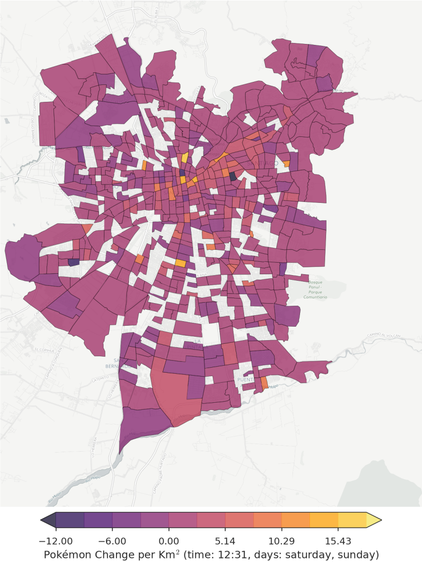


Figure 18: 12.31 weekend day

XDR: PoGo: Results (geoloc'ed)

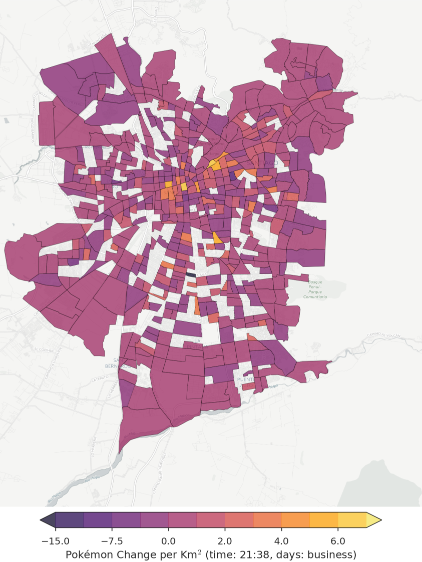


Figure 19: 21.38 business day

XDR: PoGo: Results (geoloc'ed)

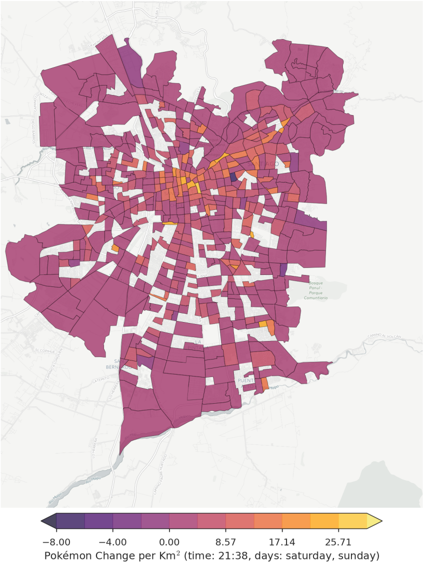


Figure 20: 21.38 weekend day

XDR: PoGo: Conclusions

- Daytime: 13% more connections,
- Night time: 10% more connection, so more people went out (at night, to parks!), making it safer for other people too!

XDR: PoGo: Conclusions

In any case, TAKE HOME:

- XDR datasets can be (very) sensistive to short-time/(lived?) events

XDR: Malls: Intro

- choice of mall is distance-based @deSimone2016
- Intuition: in SCL, malls considered social “melting pots”

XDR: Malls: Intro

- choice of mall is distance-based @deSimone2016
- Intuition: in SCL, malls considered social "melting pots"
- So, **Q**: At equivalent distances, whould you choose the more diverse mall?

XDR: Malls

- Data:
  - August 2016 XDRs
  - 16 malls in Santiago
  - 481 indoors towers
  - **387,000** individuals and **1.4M** mall visits

Spatial segregation

Loufs et al.'s definition of segregration[6]

$$E_{\alpha\beta} = \frac{1}{N_\alpha} \sum_{m \in M} n_\alpha(m) r_\beta(m)$$

where

$$r_\beta(m) = \frac{n_\beta(m)/N_\beta(m)}{n(m)/N}$$

So, intuitively, if $E_{\alpha\beta} > 1$, then mixing happens; else, segregation.

[6]Louf R, Barthelemy M (2016) Patterns of residential segregation. PLoS ONE 11(6):0157476, Link: `https://bit.ly/2Jbzthw`

XDR: Malls: HDI Segregation

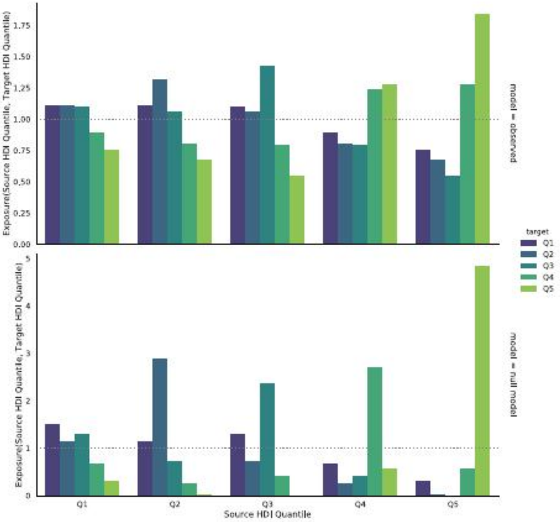


Figure 21: Results of the social mixing index, using the segregation model for observed and null models (visiting only nearby malls)

Take out and a question

- So: there's diversity... BUT: would people choose to go to a more mixed mall?

XDR: Malls: Gravity model of visits

Testing the factors that influence mall visits using Gravity Model:

$$F_{ij} = G \frac{M_i^{\alpha} M_j^{\beta}}{D_{ij}^{\gamma / S_j}}$$

where:

- F_{ij} is number of visitors from comuna i to mall j
- M_i is the population of comuna i
- M_j is the size of mall j
- D_{ij} is the distance between comuna i and mall j, and in particular
- S_j **is the diversity of mall** j, so malls with higher entropy appear as closer:

$$S_j = -\sum_{q \in Q} p_q \lg p_q$$

where p_q is the fraction of visitors to mall j that belong to HDI percentile q.

XDR: Malls: Fitting

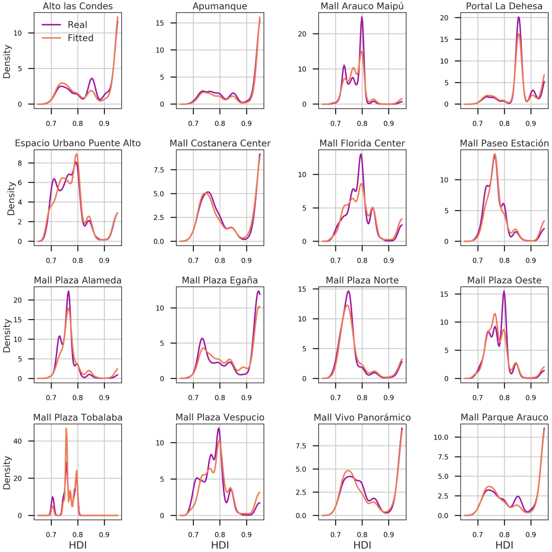


Figure 22: All coefficients positive: negative effect of distance, a positive effect of diversity on mall election

XDR: Malls: Conclusions

- After fitting the model, it is possible to predict the social characteristics of mall visitors,
- *D*istance is the most important factor when choosing malls,
- Mixing is (statistically) larger than we'd find if visitors were to visit the nearest mall. Thus,
- A positive effect of social mixing in choosing what mall to visit (at equal D and M, people prefer diverse malls).

XDR: Malls: Conclusions

In any case, TAKE HOME:

- Our data is sensitive to who goes *into* malls

DPI: News: Intro

- Vilella, Ferres, Paolotti, Ruffo (under review). **Inspecting urban inequalities in information-seeking behaviour**[7].

Q: Does reading grow linearly with HDI?

[7] Elejalde, Ferres, Schifanella. (2019). **Understanding News Outlets' Audience-Targeting Patterns**. EPJ Data Science, 8 (16) (Springer)

DPI: News: Datasets

- DPI:
  - July 2016
  - IP addresses of 27 news media outlets, for most of which we know their political alignment and ownership structure

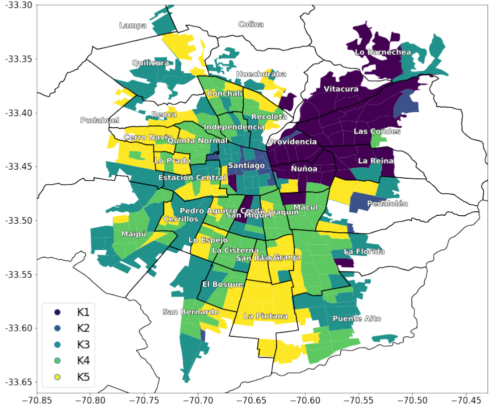
	antenna	date	hour	ip	usrs
1	00000000	20160706	11	200.12.26.117	1
2	00000000	20160706	14	190.153.242.131	1
3	00000000	20160706	14	200.12.20.11	1
4	00000000	20160706	15	190.110.123.219	1
...	...	...	...	...	...

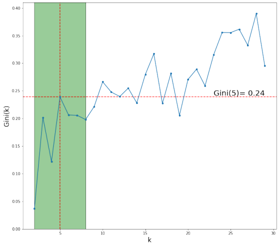
- The 2017 census (17m people, blocks)

DPI: News: Outlets

BioBioChile
El Mercurio editorial group
Cooperativa
AdnRadioChile
The Clinic
Tele 13
Publimetro Chile
Diario Financiero

DPI: News: Clustering census districts





Cluster	Mean age	Avg years of schooling	% of students	% of people of indigenous ethnicity
K1	46.25	16.91	0.15	0.05
K2	38.78	16.50	0.18	0.07
K3	42.05	14.65	0.14	0.10
K4	46.36	14.30	0.12	0.10
K5	44.62	12.86	0.11	0.13

DPI: News: General results

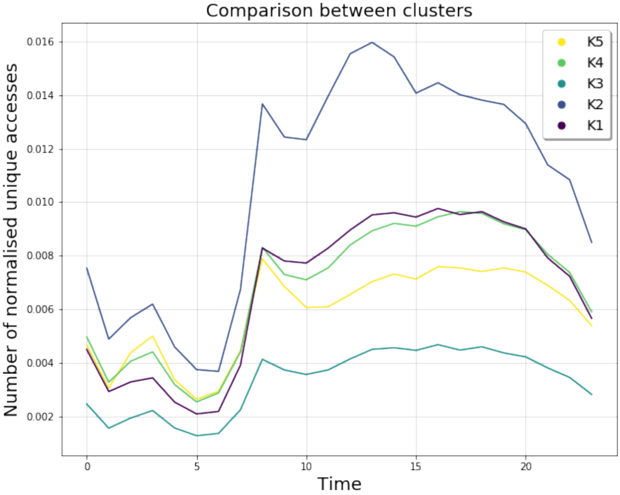


Figure 23: Young and educated read significantly more than other groups

DPI: News: General results

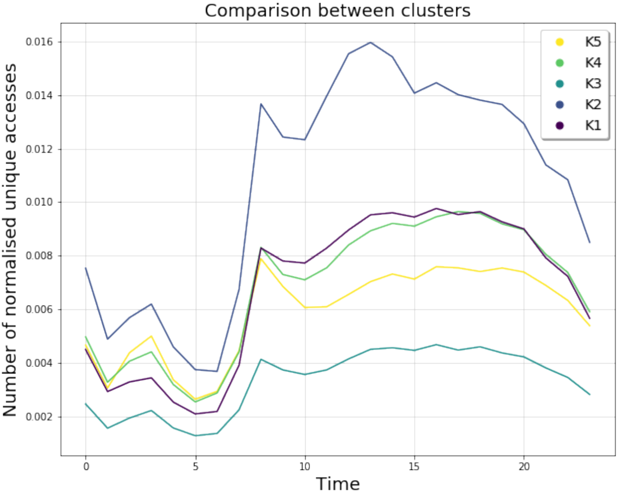


Figure 24: K3 more educated than K4, K5 (lot more!), but read less

DPI: News: Specific results

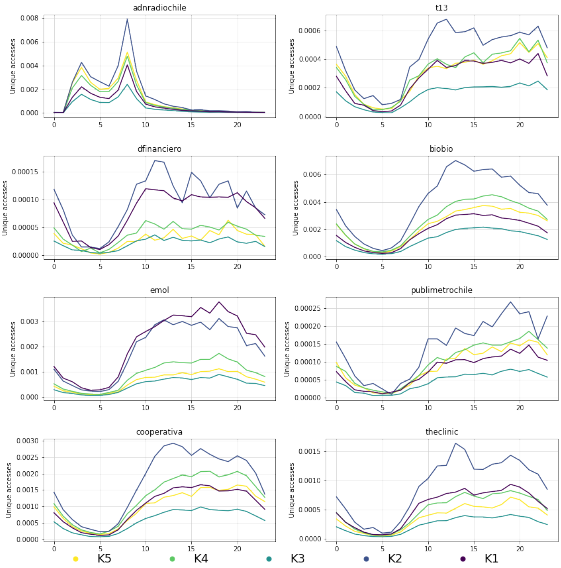


Figure 25: Young and educated read more varied content

DPI: News: Specific results

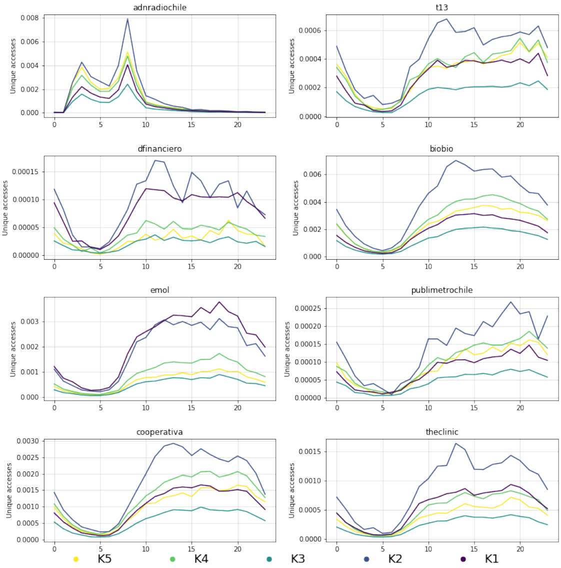


Figure 26: K1 is restricted to particular (conservative, !capitalist) outlets

DPI: News: Specific results

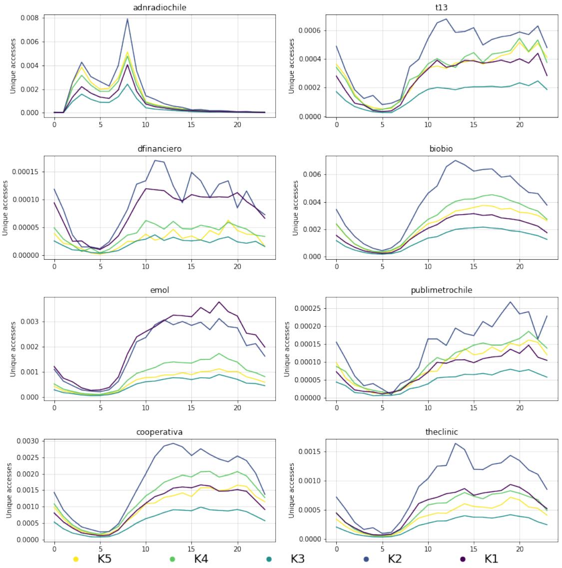


Figure 27: K3 still at the bottom

DPI: News: General conclusions

- Linearity between K does not hold

DPI: News: General conclusions

In any case, TAKE HOME:

- Even highly anonymized data can tell you a lot about information consumption

Why were the TAKE HOMES so wacky?!

- All along you learned about:
  - gender
  - social mixing
  - land use
  - news consumption

Why were the TAKE HOMES so wacky?!

- This was, all along, a story about data.

But. . .

- We know **very** little about the formal properties of these wild datasets

Thank you

Figure 28: SCL

Collaborators

Loreto Bravo (IDS, UDD & Telefonica), Eduardo Graells (IDS, UDD & Telefonica), Diego Caro (IDS, UDD & Telefonica), Daniela Opitz (IDS, UDD & Telefonica), Fran Varela (IDS, UDD & Telefonica), Pablo García (BCI), Eric Ancelovici (Telefónica), Manuel Sacasa (Telefónica), Andrés Leiva (Telefónica), Ciro Cattuto (ISI Foundation), Daniela Paolotti (ISI Foundation), Laetitia Gauvin (ISI Foundation), Michele Tizzoni (ISI Foundation), Johan Bollen (Indiana University), Rossano Schifanella (U Torino), Giancarlo Ruffo (U Torino), Erick Elejalde (L3S, Germany), Markus Strohmeier (Aachen, Germany), Eelco Herder (Radboud, The Netherlands), Bruno Goncalves (JP Morgan, USA), Stefaan Verhulst (NYU, USA), Natalia Adler (UNICEF, USA), Ricardo Baeza-Yates (Northeastern@Silicon Valley), Salvatore Vilella (ISI, UTorino), Meng He (Dalhousie, Canada), Travis Gagie (Dalhousie, Canada), Norbert Zhe (Dalhousie, Canada), Mariano Beiró (UBA, Argentina), André Panisson (ISI Foundation), Michel Dumontier (Maastricht, The Netherlands), Karim Touma (Falabella)