Unveiling Chagas with Big Data

Carlos Sarraute[1] Carolina Lang[1,2] Alejo Salles[2] Juan de Monasterio[2] Diego Weinberg[3]

[1]GranData Labs

[2]Universidad de Buenos Aires

[3]Fundación Mundo Sano

Workshop: Big Data and Environment
October 12th 2015

Agenda

Presentation

Grandata

- Started in 2012.
- Labs team: six people based in Vicente Lopez (Buenos Aires).
- Researching Human Dynamics.
  - Using "Big Data" to analyze social networks and human behavior.
  - Integrating banking and cellphone data.
  - We aim to characterize and predict user actions.

Presentation

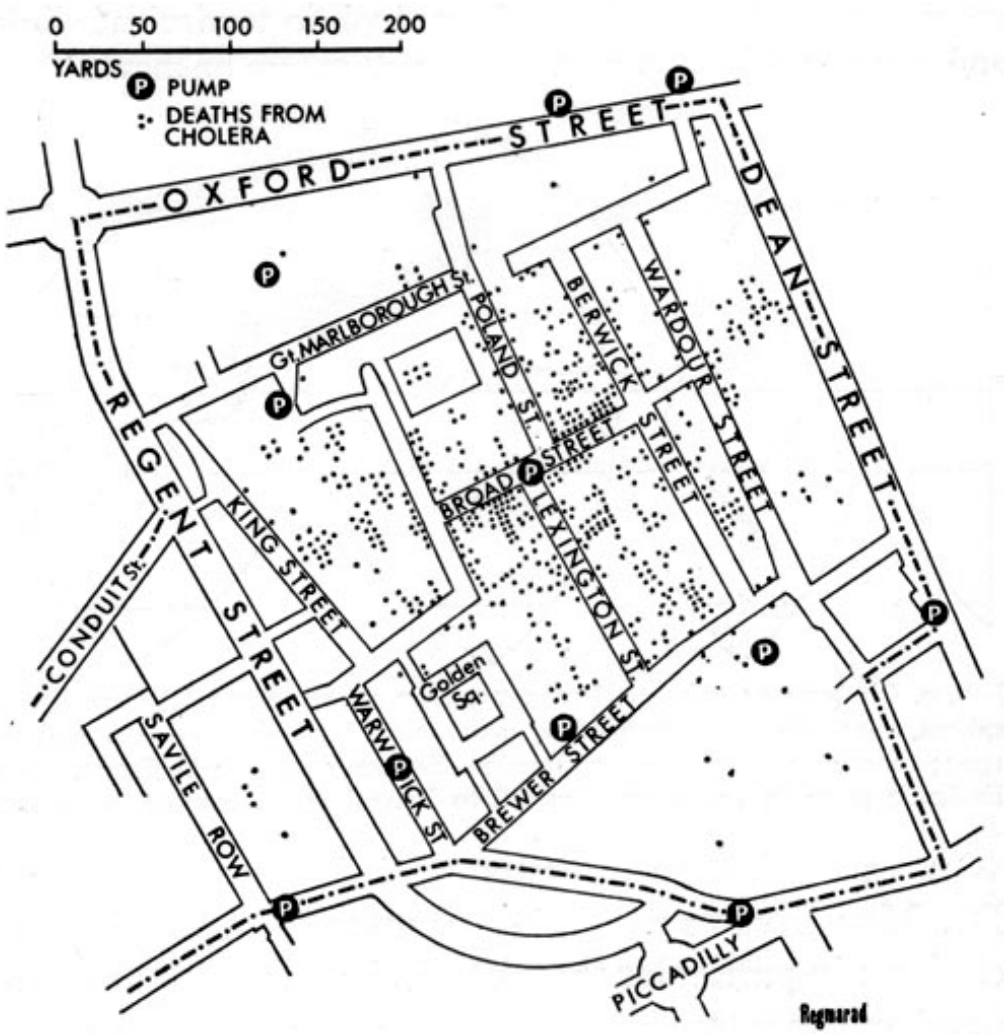
Grandata

- Started in 2012.
- Labs team: six people based in Vicente Lopez (Buenos Aires).
- Researching Human Dynamics.
  - Using "Big Data" to analyze social networks and human behavior.
  - Integrating banking and cellphone data.
  - We aim to characterize and predict user actions.

Scientific Collaborations

- Aline Viana (Inria, Paris)
- Eric Fleury, Marton Karsai (ENS,Lyon)
- Sandy Pentland and the Human Dynamics Lab (MIT)
- UBA and Mundo Sano Foundation

Data Visualization & Epidemics



London cholera cases.
Map made in 1854 by Doctor John Snow.

Mobile Phone Data

Our dataset ranges 5 months of **anonymized** call detail records (CDRs) from one national telco.

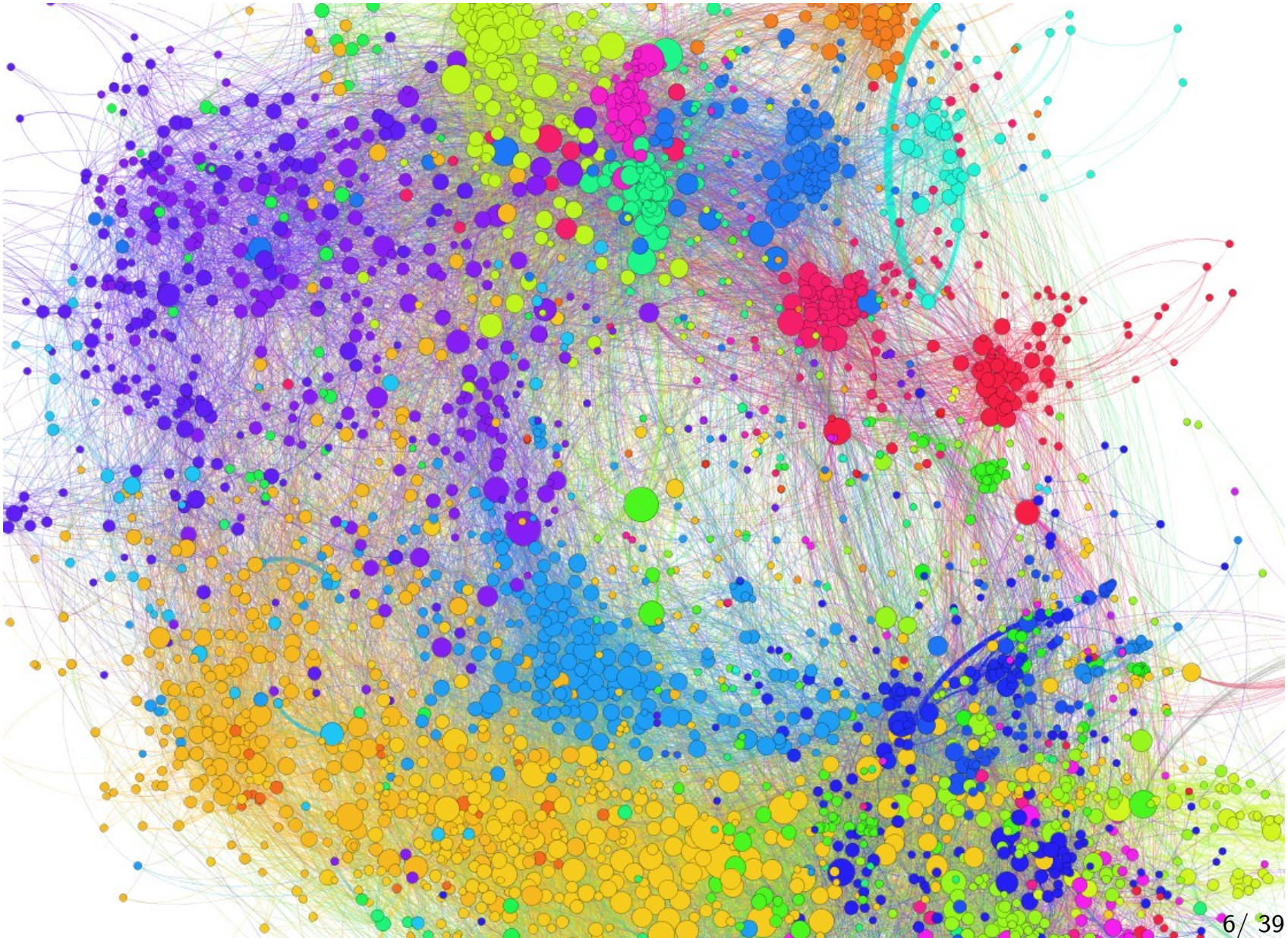
Each call record consists of:

- Origin and destination users.
- ID of origin antenna.
- Start time and duration.

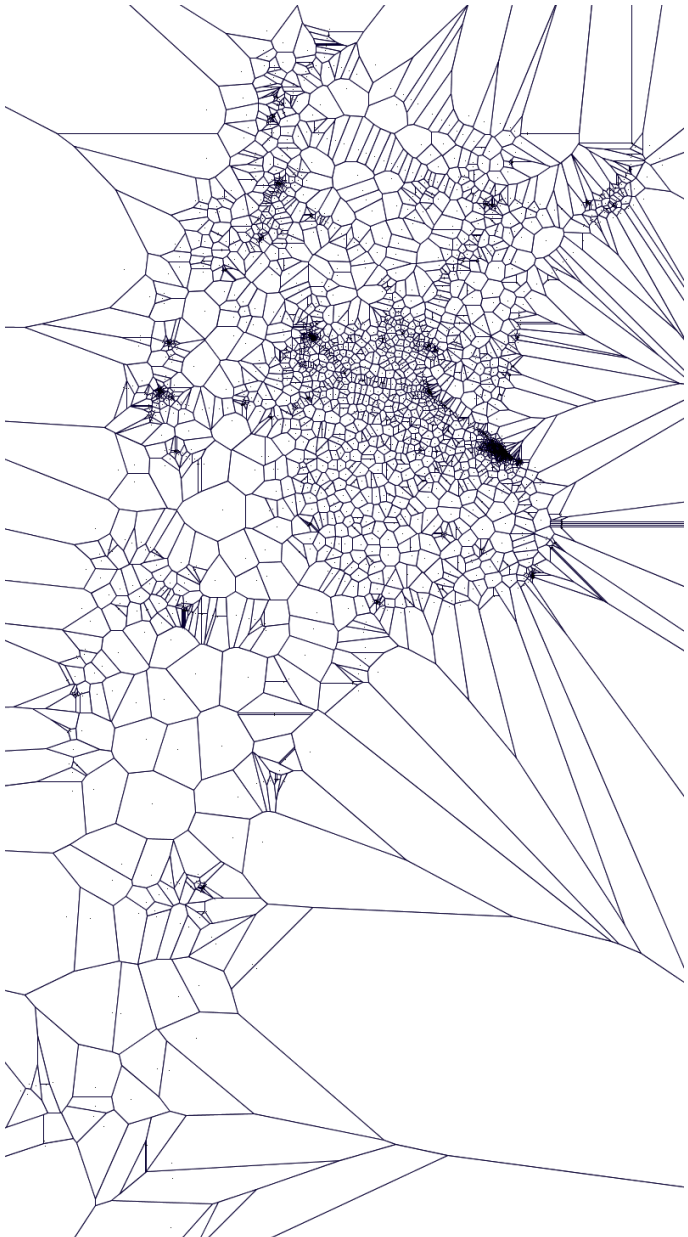
In all, there are more than 9 billion geolocalized calls from 40 million mobile lines.

The antenna dataset consists of more than four thousand geolocalized antennas.

Social visualization: The Communications Graph



Spatial Visualization: Voronoi Cells from Antennas



Agenda

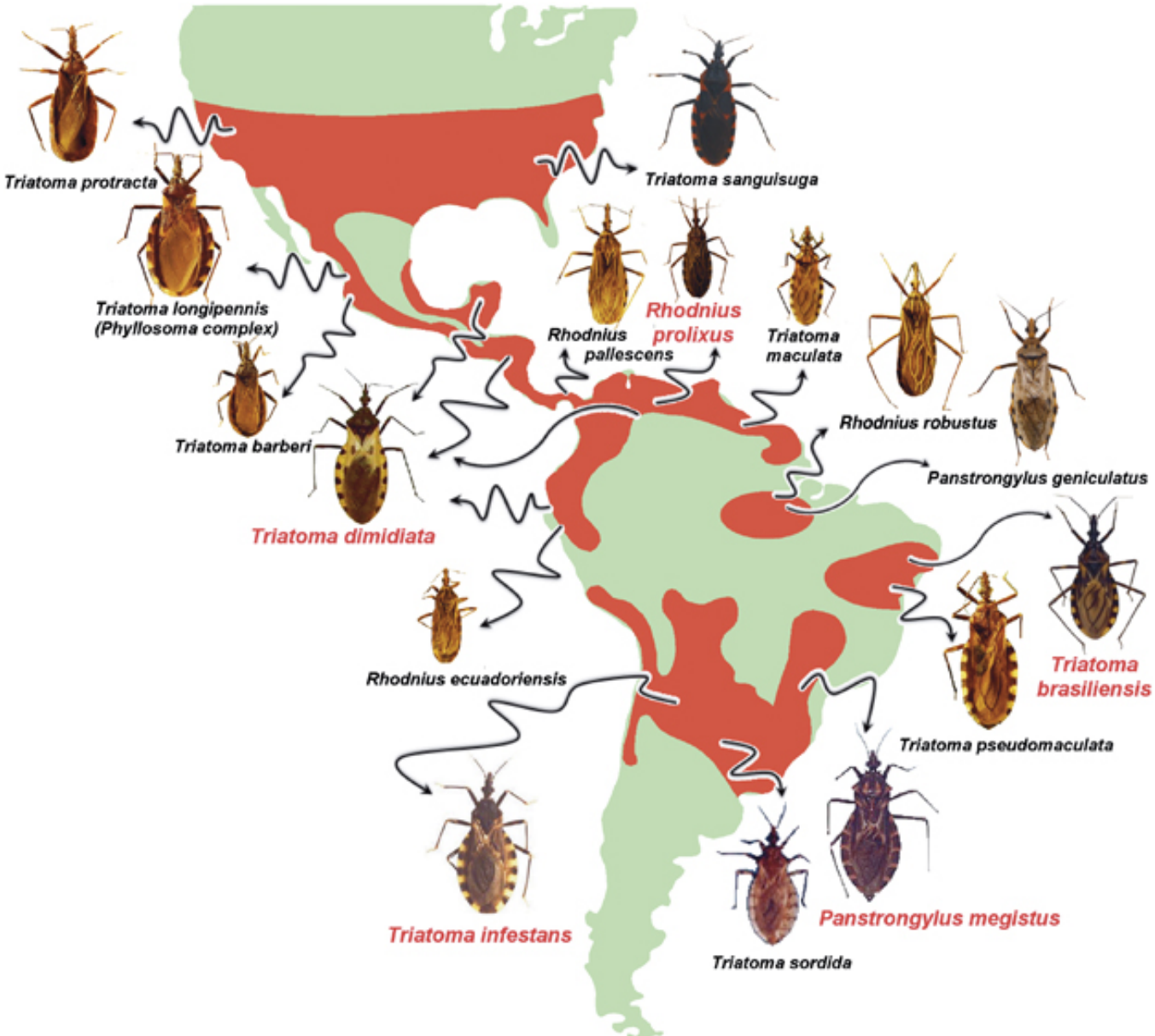
Epidemiology of Chagas

Chagas is a disease caused by the *trypanozoma cruzi* parasite that covers the whole American continent and some parts of Europe.

With over 65 million people exposed and endemic in more than 21 Latin American countries.

WHO estimates between 6 and 8 million infected individuals worldwide of which only 1 % have access to diagnosis and treatment.

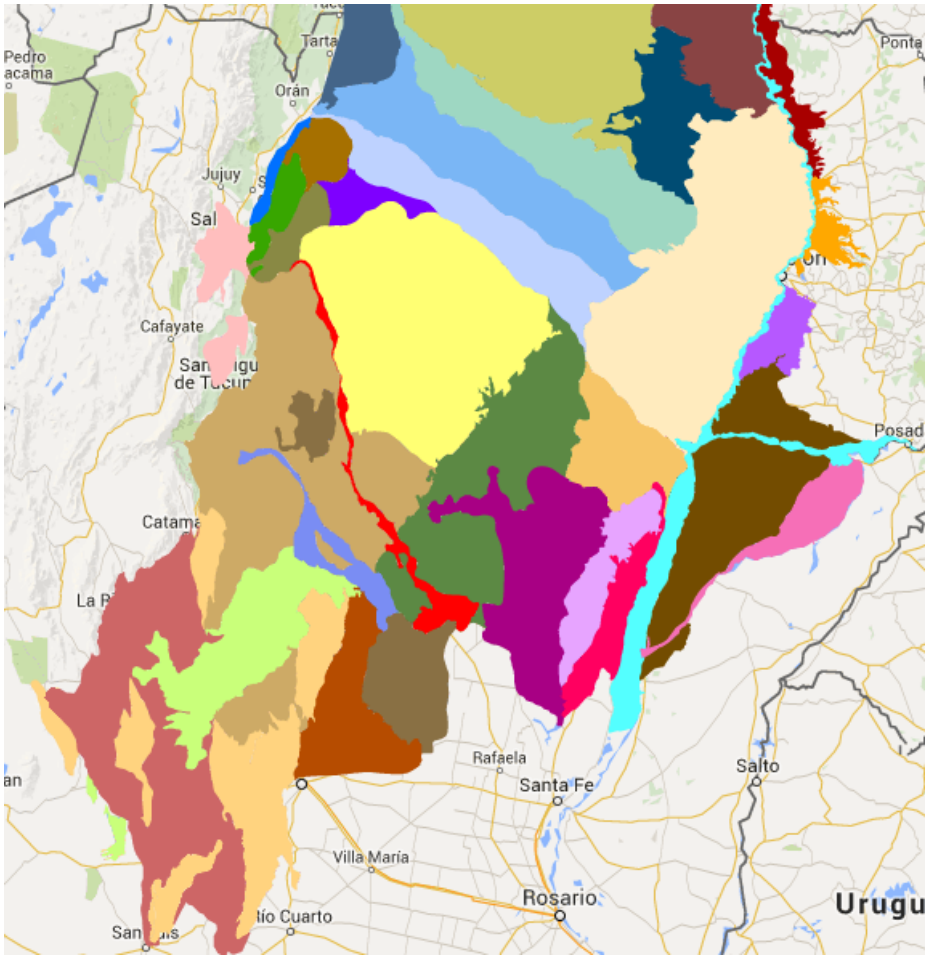
Epidemiology of Chagas



Epidemiology of Chagas

In Argentina, the disease is endemic in the *Gran Chaco* region and also predominant in communities with migrations from this area.

National estimates account for at least 1.5 million infected users and only 1 to 2 treatments done yearly. 30 % of all infected will develop a cardiopathy.



Epidemiology of Chagas

Chagas is a disease:

- ... which is spread mostly through its vector (the triatomine bug) and congenitally from mother to baby. In smaller measure, transmission also exists by blood transfusion and organ donations.
- ... whose asymptomatic phase can last more than 10 years.

We aim to find infected people living in areas which are traditionally non-endemic. Living outside *Gran Chaco* and carrying the disease parasite *trypanosoma cruzi* that causes Chagas.

Objective: Chagas and Migrations

Our goal is to find those places which have most of the migratory exchange with the ecoregion.

Having a disease with long asymptomatic phases imply that long-term migrations are relevant to analyze.

The goal is to find...

- individuals which have been infected in endemic areas and have later migrated.
- seasonal workers whose residence varies during the year.
- outlying communities with probability of having a high Chagas prevalence.

Methodology (1)

Home Prediction

- As a first step, we determined each user's residence antenna.
- This was chosen to be the most used frequently used antenna, considering only calls made in week evenings.
- The hypothesis: most of people are at home on any given weeknight.
- Note: users for which the inferred home antenna is located in *Gran Chaco* (the risk area) will be considered the set of *residents of Gran Chaco*.

Methodology (2)

Vulnerable users

- For every user, we listed all of the call receivers in a given month.
- If a given user communicated with the *Gran Chaco*, we considered him to have higher risk of having Chagas and we tagged him as potentially *vulnerable*.

Methodology (2)

Vulnerable users

- For every user, we listed all of the call receivers in a given month.
- If a given user communicated with the *Gran Chaco*, we considered him to have higher risk of having Chagas and we tagged him as potentially *vulnerable*.

Antenna aggregation

- We aggregated vulnerable users and total users (residents) per antenna.
- We also aggregated the total volume of outgoing calls from every antenna and from these we extracted every call that had a user whose home is in the *Gran Chaco* area as a receiver (*vulnerable calls*).

Methodology (3)

Heat Maps

We generated heat maps from the processed data where each cell was represented by its communications with the ecoregion.
We generated a circle for each cell where:

- the **area** depends on the on the volume of use.
- the **color** corresponds to the percentage of use which is vulnerable in that antenna.

Methodology (3)

Heat Maps

We generated heat maps from the processed data where each cell was represented by its communications with the ecoregion.
We generated a circle for each cell where:

- the **area** depends on the on the volume of use.
- the **color** corresponds to the percentage of use which is vulnerable in that antenna.

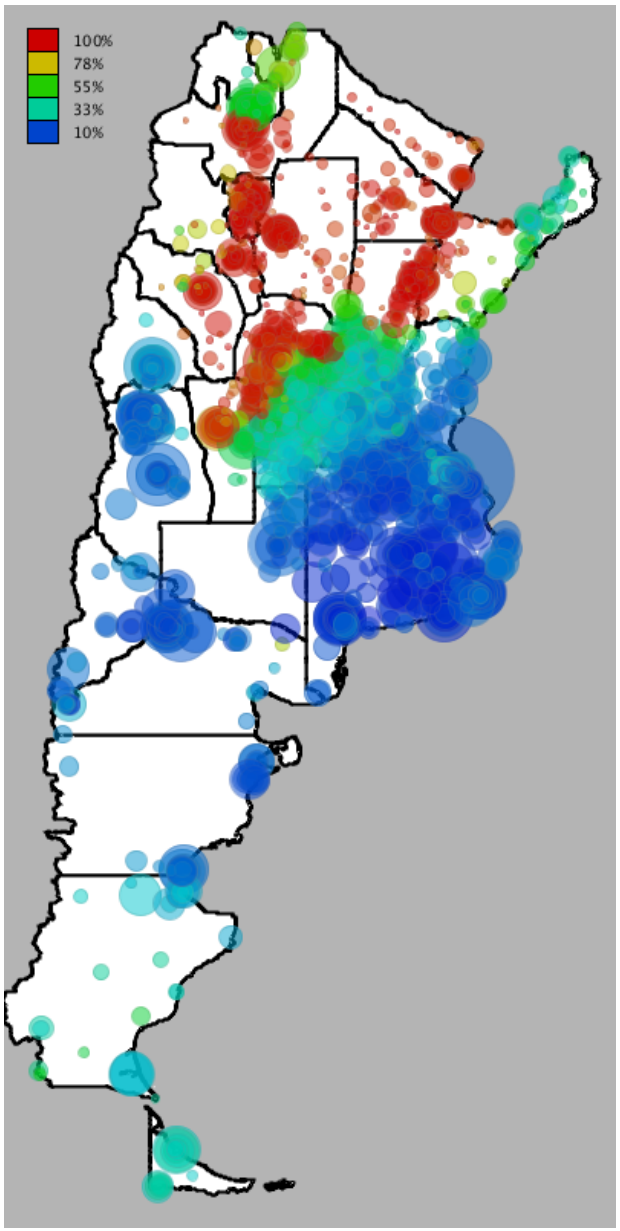
Antenna Filter

We built two filtering parameters β and $min\_volumen$ which will control the antennas to be plotted. Every antenna will be plotted if:

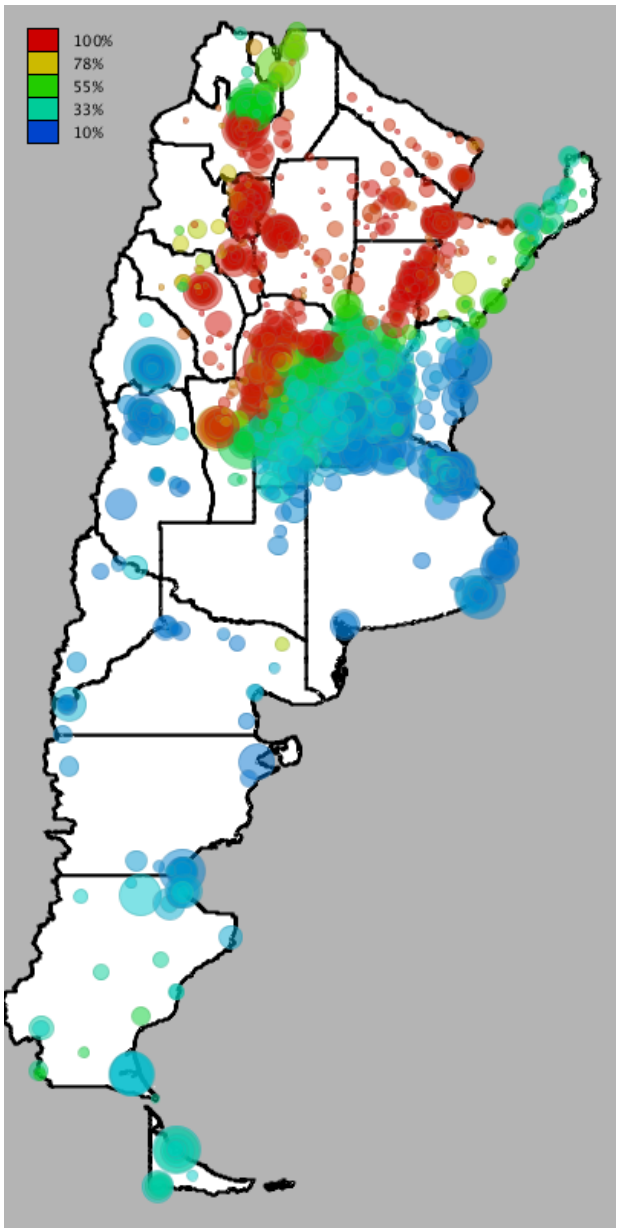
- the percentage of vulnerable users/calls is bigger than β.
- the volume of vulnerable users/calls is bigger than $min\_volumen$.

Agenda

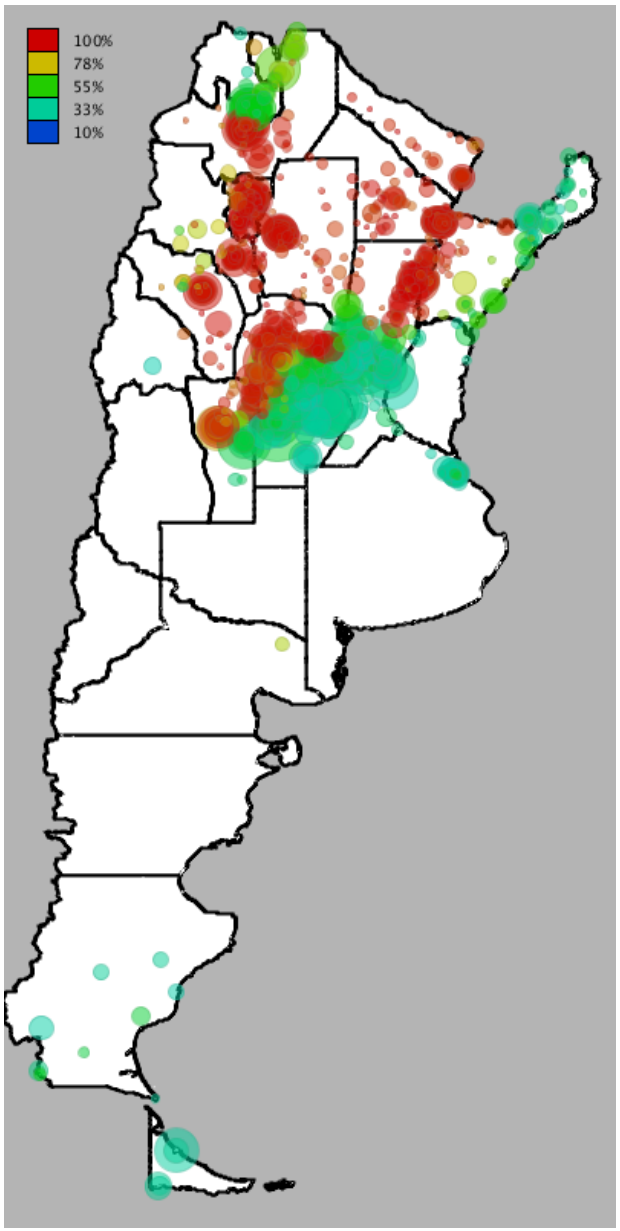
Heat Map Argentina, $\beta = 1\,\%$



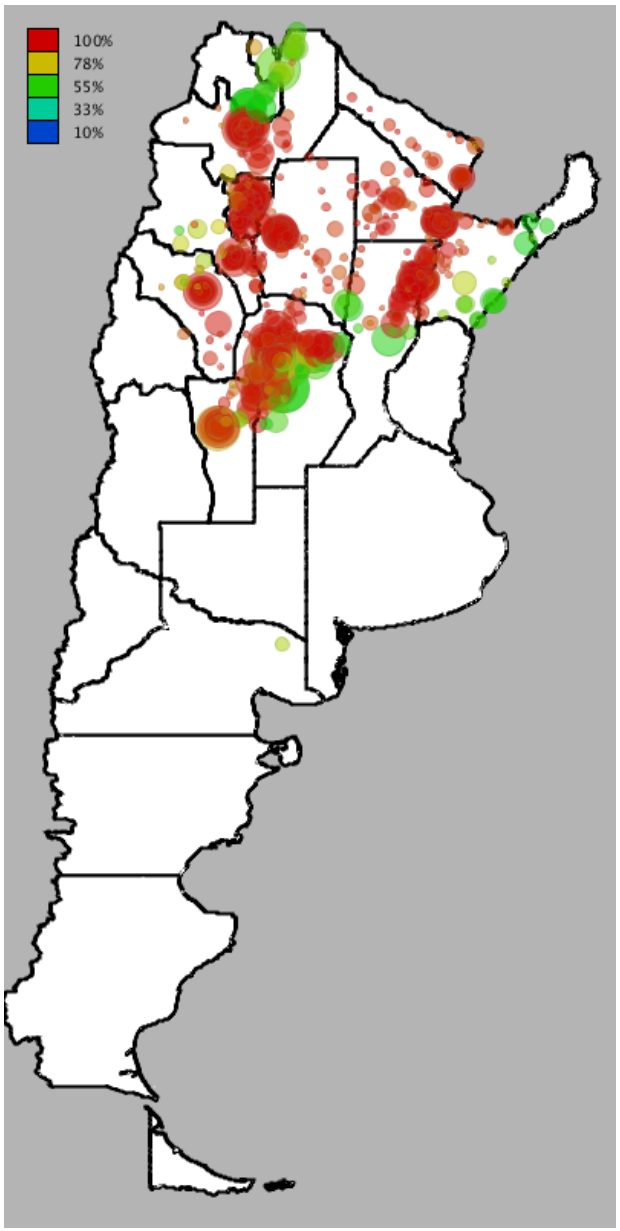
Heat Map Argentina, $\beta = 15\,\%$



Heat Map Argentina, $\beta = 30\,\%$



Heat Map Argentina, $\beta = 50\,\%$



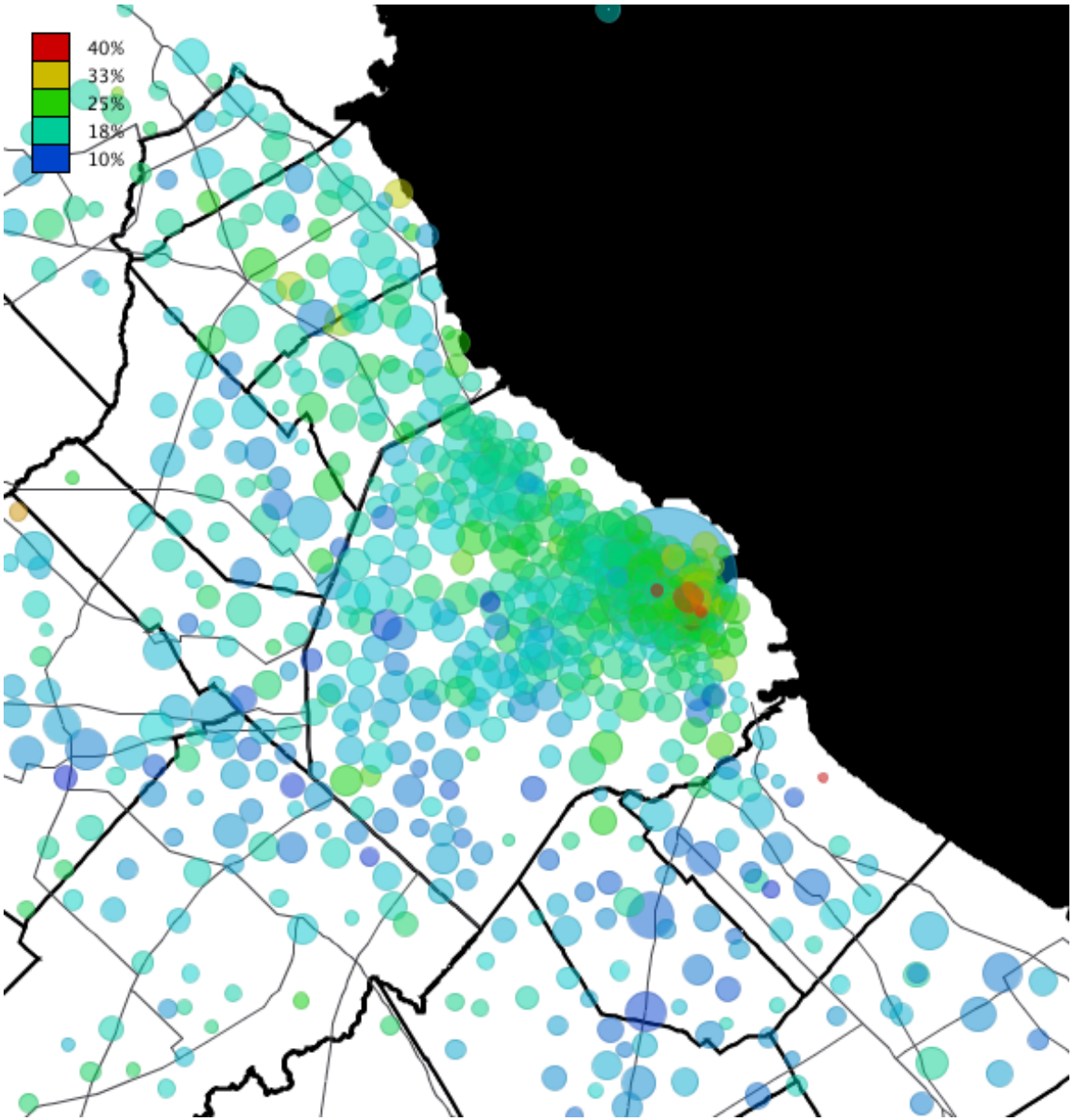
Focused Areas

From these first maps, we have decided to focus visualizations to enhance precision in certain regions outside of Gran Chaco.

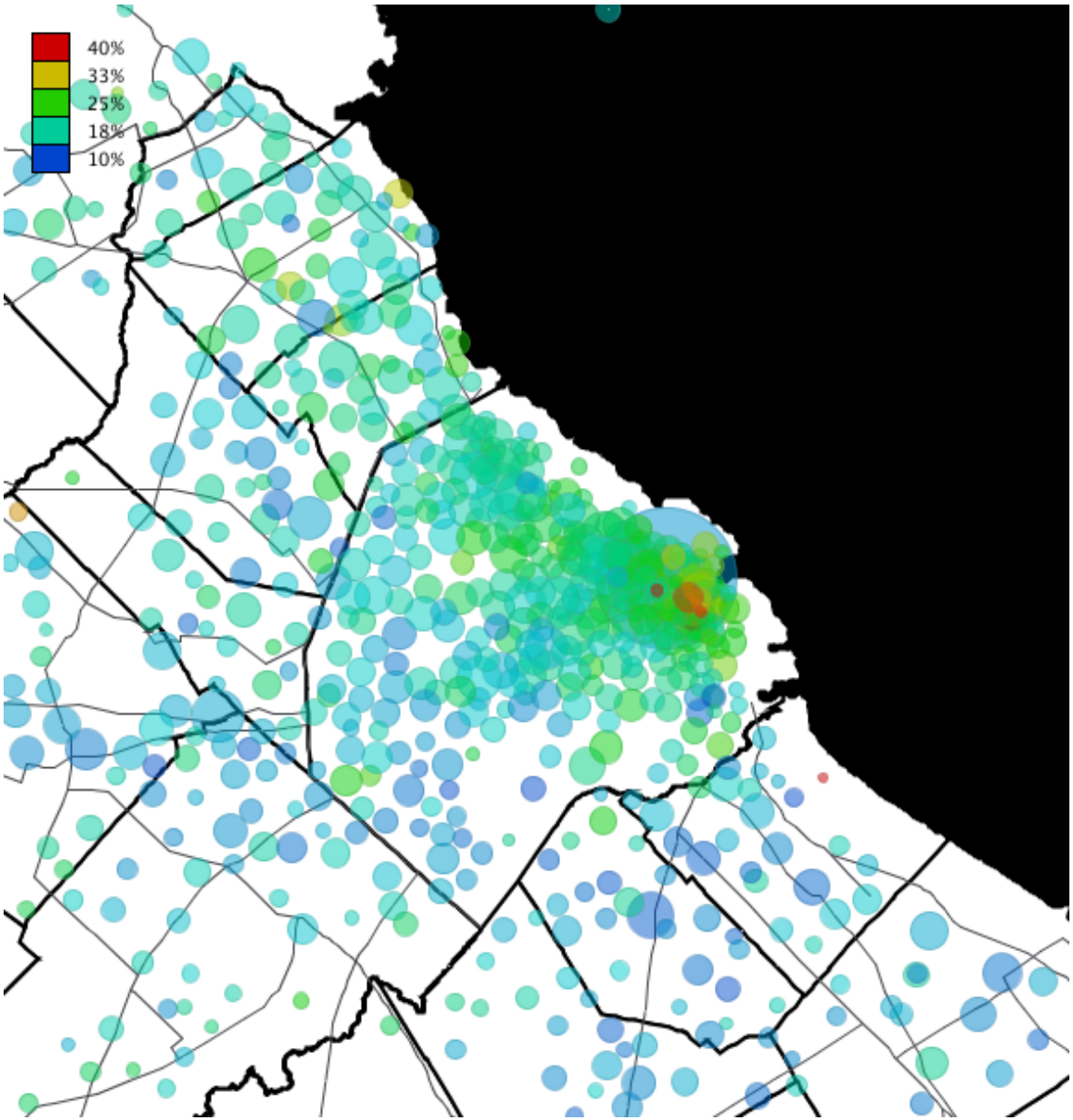
Predetermined Areas

- Tierra del Fuego and South Santa Cruz
- Chubut.
- East Río Negro.
- Buenos Aires.
- Capital Federal, South and North CABA, and AMBA.
- Central Córdoba and Santa Fe.

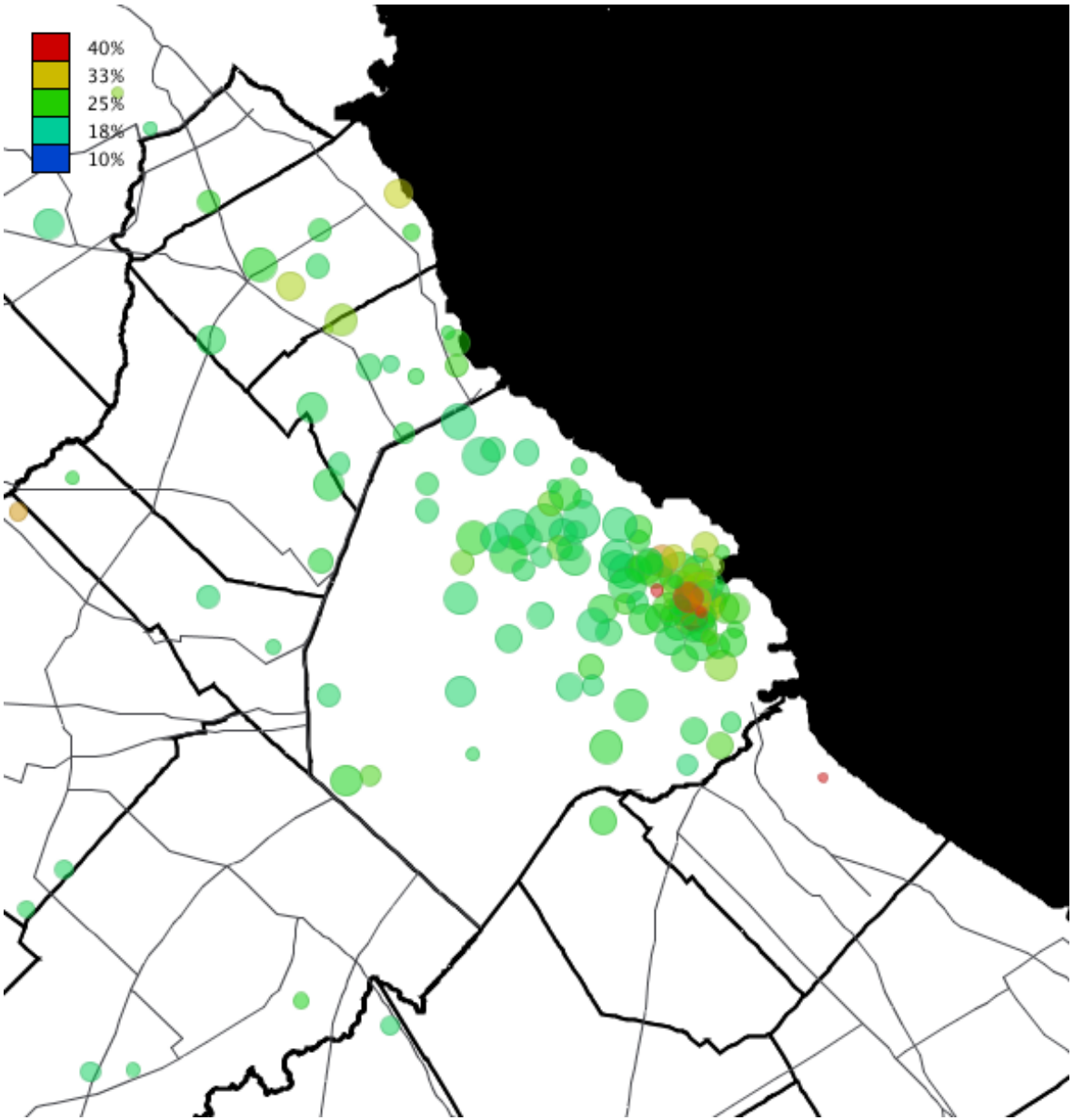
Heat Map AMBA, $\beta = 2\,\%$



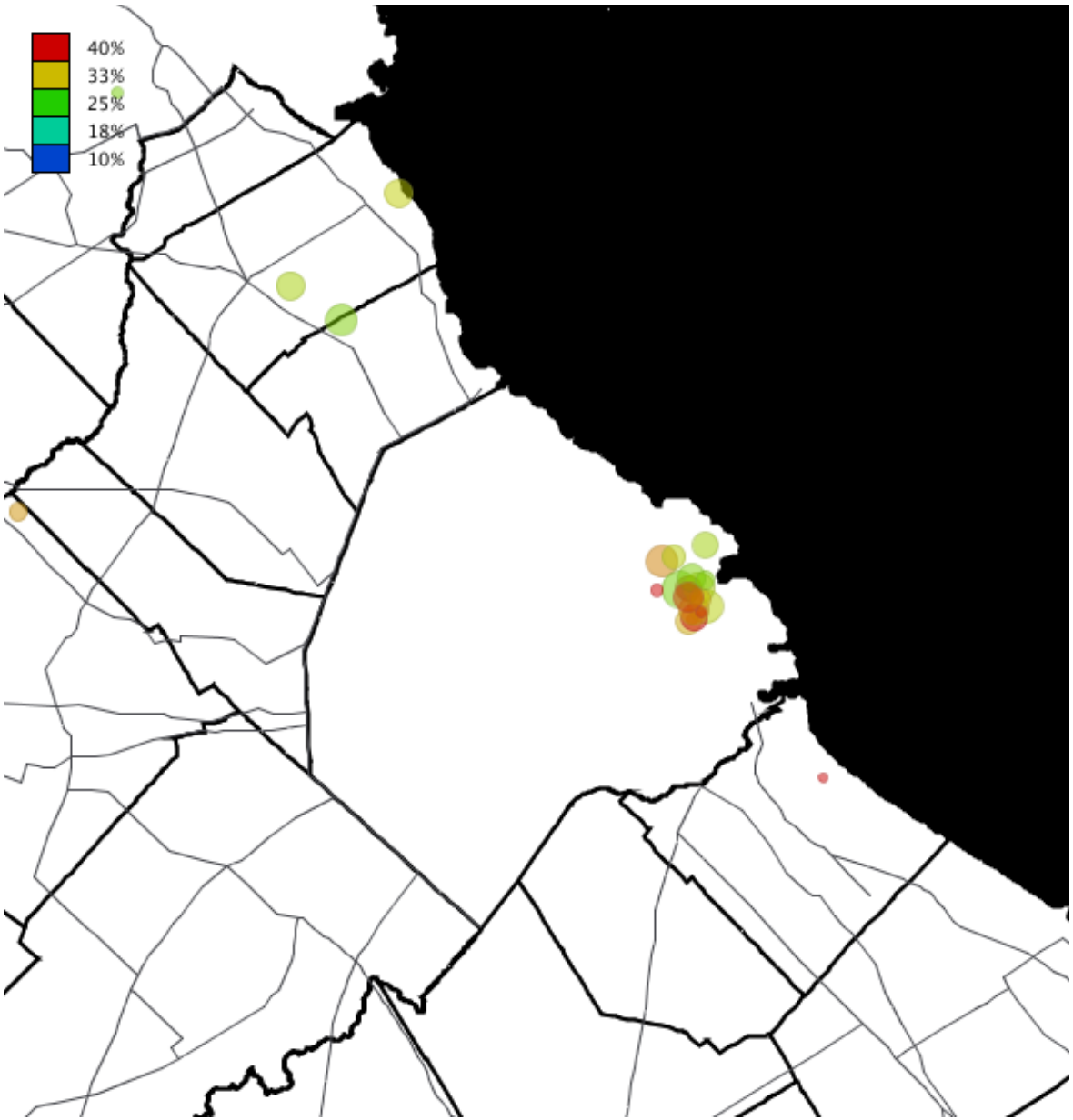
Heat Map AMBA, $\beta = 10\,\%$



Heat Map AMBA, $\beta = 20\,\%$



Heat Map AMBA, $\beta = 28\,\%$



Communities Detected

After filtering the maps, we are left with few antennas that stick out for their risk level. Using their location, we assign them to Argentinian communities or cities.

Some communities outside the big cities

- Cordoba: Freyre, La Tordilla, Balnearia, ...
- AMBA: Avellaneda, Parque Patricios, San Isidro, ...
- Bs As Province: Lima, San Nicolas.
- La Rioja: Chamical and Malanzán.
- Salta: Tartagal.

Ongoing Tasks

Work vs. No Work

Anatuya and Pampa del Indio

Agenda

Conclusions

Expected

Heat maps show temperature falling from **Gran Chaco** outwards; heat that descents rapidly as we move away from the area.

Conclusions

Expected

Heat maps show temperature falling from **Gran Chaco** outwards; heat that descents rapidly as we move away from the area.

Unexpected

We also find outlying communities that stick out for their high link with the studied region, way over other communication patterns of the area where they are located. This confirms that communication patterns with the ecoregion are inhomogeneous.

The maps provide a tool to **prioritize screening campaigns for Chagas Disease**.

Conclusions

CDR Reuse

- A novel use for CDRs is presented, data which is originally logged for other purposes (billing).
- The dataset analysis is of low-cost and potentially of hight-value.
- This proof of concept can be extended to other regions or diseases and epidemics of similar characteristics.

Future Work

Enhance the Communications Model

- Current vulnerability characterization relies exclusively on mobile communications of users with antennas in the ecoregion.
- We look forward to extracting long-term mobility information by looking at communication patterns between two areas.

Future Work

New Data Sources

With Mundo Sano we aim to incorporate epidemiological data to the analysis:

- tenement bug infestation
- infected rates by municipality
- infected newborns
- acute Chagas cases
- serological data per community, amongst others.

2010 Census

Socio-economic information aggregated by department from the national census website [a] has already been preprocessed for every municipality in Argentina.

[a] http://censo2010.indec.gov.ar/

Acknowledgements

¡Thank you!

- .. Grandata
- .. Mundo Sano Foundation
- .. to you!

¿Questions?

Carolina Lang
carolang@grandata.com
Juan de Monasterio
laterio@gmail.com