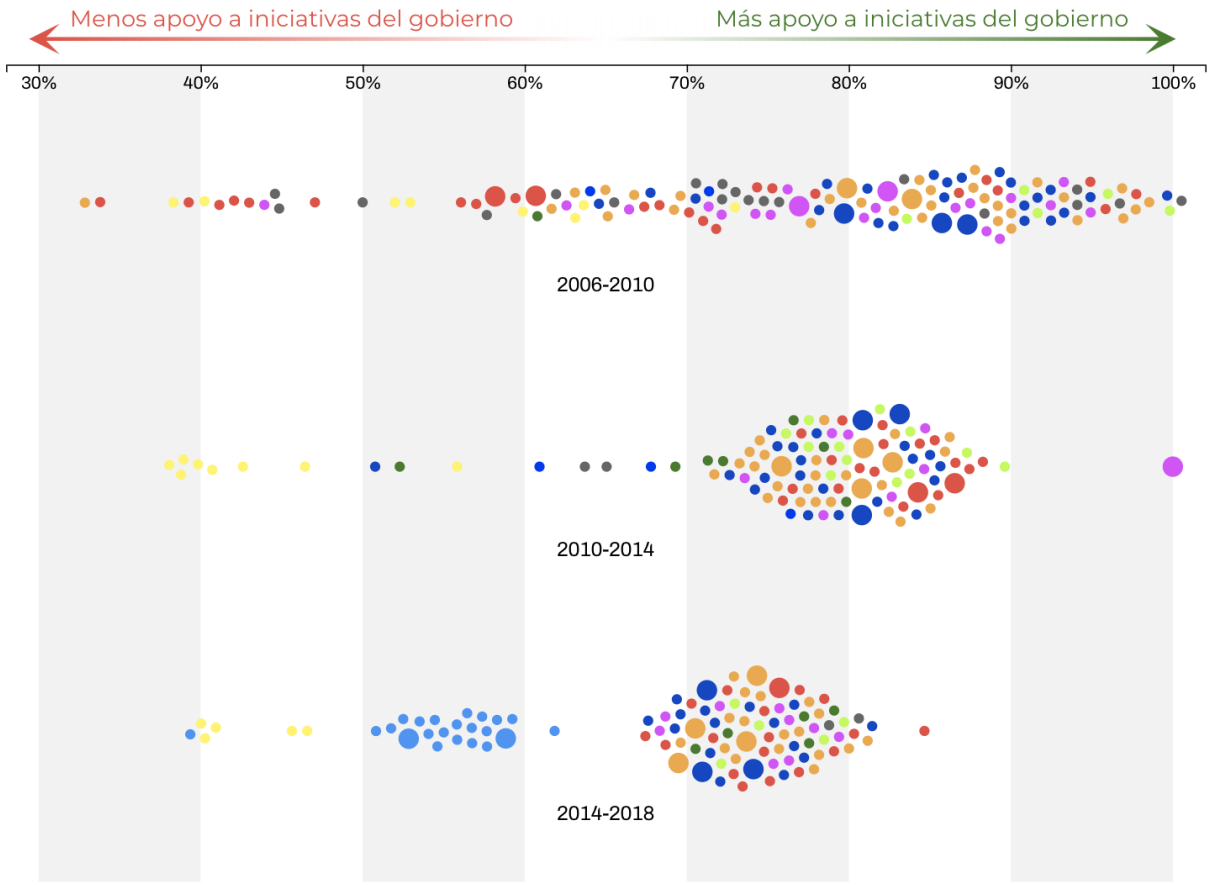
Scrollytelling: Visualizing voting patterns across members of the Colombian congress

Juan J. Castro and John A. Guerra



Abstract—Politics has an intrinsic accountability problem: officials are elected to represent the interest of their constituents, but the latter cant be aware with the particulars of the day to day actions of the former, particularly when it involves thousands of votes. This creates a knowledge imbalance that makes political accountability and transparency harder. Congreso Visible (CV) is an organization dedicated to make the day-to-day actions of Congress more visible and accountable. Together weve worked to identify, highlight and visualize insights by turning CVs vast data silos containing years of voting data into a web-based storytelling experience using data visualization.

1 Introduction

As the world becomes more connected, our interactions more digital and our work is increasingly mediated by computers, we're collecting more data than ever before. Organizations of all kinds collect data to track and achieve goals, generate insights and evaluate performance metrics, among other applications. Data visualization is a key tool in most of these tasks, since it works to efficiently summarize large amounts of quantitative information into a digestible, graphical package.

Visualizations of large amounts of political -or otherwise "social"- data have become popular in academia, news publications and governmental think tanks, usually backed by a larger trend of open data: every day more organizations and governments are opening up the doors to their untapped

- *J. A. Guerra is with Universidad de los Andes, UCBerkeley. E-mail: john.guerra@gmail.com*
- *Juan J. Castro is with Universidad de los Andes. E-mail: jj.castro10@uniandes.edu.co*

Manuscript received XX, 20XX; revised XX, 20XX.

data resources, usually in the interest of transparency and openness.

Today, working within these constraints, several teams exist like the New York Times Graphics Department (NYT Graphics), entirely dedicated to generating open, easy-to-understand visualizations to support their journalistic endeavours. These teams take advantage of the speed at which data can be collected, summarized and aggregated nowadays to create stories backed by data visualization, something that has never been possible before in the realm of journalism.

Congreso Visible (CV) is an organization based out of Universidad de los Andes dedicated to making the day-to-day actions of Congress more visible and accountable. CV has dedicated a large amount of work towards turning the daily publications from the Colombian Congress into public datasets. These publications, formally called gacetas, which include such information like voter data and law project metadata, are often hard to parse and gain information from, since Congress does little to summarize or process this information.

For this project, working within the Visual Computing research group IMAGINE, weve explored the possibilities of turning CV's data silos into data visualizations to support CV's insights and publications. We've created a set of data views and their corresponding visualizations focused on summarizing the data into a single web-first story.

Working alongside the team at CV, the team worked to create a first prototype which focuses on pro- and anti-government voting patterns in of congress. This document focuses on how the team created this prototype, the context in which it was created and the lessons learned from the process.

2 Political context

Colombian politics has changed significantly in the last 30+ years. What was a historically bipartisan landscape, dominated by the Liberal and Conservative parties, has become a multiparty system with dozens of parties occupying seats in the legislature. This was enabled by a complete rewrite of the Constitution of Colombia which took place in 1991. Embedded in this governing document of the republic, it was now much easier to create and elect smaller political parties, as well as new incentives and new financial constraints to limit a two-party rule. By the next decade, the number of parties grew dramatically, and began a shift to a new era of Colombian politics in which individual politicians and what they stand for carry much more weight than the party they are associated with.

As parties have come to mean less in the grand scheme of things, and outside of the many advantages of more voter choice, this has come to complicate voter decision and means that association to a single party is no longer enough: voters must be constantly up to date on the changing landscape of party and individual allegiances. lvaro Uribe, who was president of Colombia from 2002 to 2010, for example, changed parties 4 times in the years between 2001 and 2013 alone.

The last 15 years of Colombian politics have been made harder to understand by this ever changing web of allegiances. The last two ex-presidents of Colombia, lvaro Uribe and Juan Manuel Santos, both serving 8 years terms and once close allies, became fierce public rivals after the reelection of the former. As a result, once pro-government individuals have come to be anti-government in a matter of months, even when official party allegiances say otherwise. This has been further complicated by the Colombian Peace Process and subsequent agreement, which ended 60 years of armed conflict and on which both individuals stand on opposite ends of the support spectrum.

This is the context in which our work has been created. Although the news media in Colombia is widely consumed and widespread, there has never been a data-backed analysis done on this changing web of pro- and anti-government patterns within the Colombian Congress. This is partly due to the fact that, as discussed before, the data had not been available before.

3 Design and Implementation

The design and implementation process is outlined below.

3.1 Data

As outlined above, Congreso Visible's data silos contain several years of records of the day-to-day actions of the Congress of Colombia. This was the starting point for the project.

As is often the case for the nature of this data, this was stored in large Microsoft Excel files exceeding 800 MB in file size. The iterative process used to process this data consisted in several weekly cycles including some or all of the following steps:

1) **Data auditing:** Several different Tableau workspaces were used to detect anomalies and outliers in the data.
2) **Filtering outliers:** Outliers in the data were filtered to prevent misconstrued or skewed visualizations.
3) **Fixing structural errors:** As the data has been collected and changed over the years, some variables were consistently misrepresented. In most cases, this was possible to fix using a retroactive process by referring to data that is easily available on the public record, such as what representative was registered to which party in which period.
4) **Handling missing data:** For many reasons, there was missing data in some time intervals over the past 10 years. To mitigate this problem, the team decided to either extrapolate the data in some cases with reasonably long time frames, or to focus on specific time intervals in which the data was present.

3.2 Medium

As the idea of working on this project came together, the team saw the web as the clear choice, since this is the area of expertise of the participants as well as the medium in which Congreso Visible (CV) publishes data and reports. Eventually, the team saw the final product as a likely candidate for being published on the CV web portal.

After considering multiple options, we landed on Mike Bostock's D3.js [1] and Adam Pearce's graph-scroll.js as

the frameworks of choice to implement the visualization. Not incidentally, both Bostock and Pearce worked at NYT Graphics, whose publications were in great part the inspiration for the format of the finished result.

As such, the final result is a fixed visualization on the right hand side (or the top, if the site is viewed on smaller, vertical screens) that changes based on scroll events monitored by graph-scroll.js. This has the advantage of avoiding a scroll "hijacking" in which the website forces a scrolling pace or anchors to display content. Instead, graph-scroll.js monitors scrolling thresholds and triggers events based on the viewport position of explanatory text sections, so that the user can scroll at their own pace. This means that when a text section is visible on the screen, the visualization canvas is updated with the corresponding view of the data.

3.3 Design

The design of the final visualization followed an iterative design process based around weekly or biweekly prototypes and review sessions.

The final result is a scrollable set of visualizations centered on two topics.

3.3.1 First visualization set ("The votes of the Senate")

The first visualization set is a force simulation using the d3force package, the main feature of which is two beeswarm plots that highlight the differences between the last three Senate distributions and their corresponding seats and represented parties. For context, the Senate is the Upper House of the Congress of Colombia, usually with fewer representatives that the lower house.

This first set of visualizations features:

- Interactive callouts to parts of the explanatory texts, so that mouse interaction highlights explanations so specific parts of the beeswarm plots.
- A tooltip triggered by hovering on each body (each body in the plot corresponds to a different senator), to show more information.

A great deal of the groundwork for the design of the first prototype of this first set was based on J.A. Guerra's

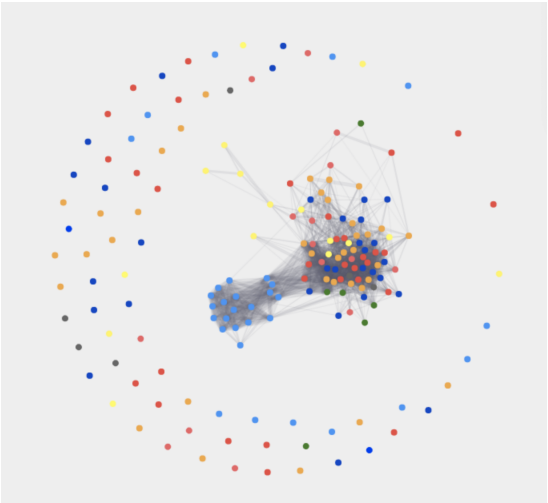


Fig. 1. Group-in-a-box algorithm applied to CV's voting data.

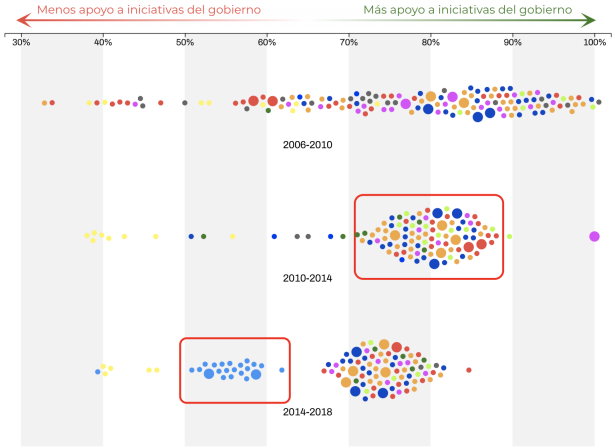


Fig. 2. Interactive callouts directing the user's eye (red) in the beeswarm plot. (The color scale is visible on the top.

prototypes on Group-in-a-box implementations using CV's data. These involved creating a d3.js force extension that implements the Group-in-a-box layout algorithm [2] to distribute nodes in a network according to their clusters. The algorithm uses a treemap to compute focis that are later used to distribute each cluster into its own box.

This visualization served to illuminate some insights that served as a starting point for some of the ideas in the project, as well as illustrate how to achieve with some technical implementation details for this first part.

3.3.2 Second visualization set ("How have the votes changed?")

The second visualization set is based around a unified line chart displaying an average of the percentage of "yes" votes over the total votes. It's an aggregated value per party, as opposed to the first set which is an aggregated value per congressperson.

The main storytelling device in this set is the continuous adding of different parties to the chart as the user progresses. As will be showed in the next section about validation, this proved to be much more user friendly than having the user interact with the graph to filter out specific parties.

This second set of visualizations features:

- Callouts in a similar style to the first set, but set in a slightly different fashion: these also highlights trends and trend directions rather than merely direct the user's eye.
- The same axis convention as the first set, but this time on the Y-axis.

3.4 Explanatory text

The team at Congreso Visible was in charge of writing and editing the texts that would go on the final site. However, since the process was iterative and involved a lot of back and forth, we needed to find a way to be able to edit this on-the-fly while still keeping it easy to integrate into the main visualization.

As part of this process, we decided to not reinvent the wheel by bootstrapping or creating a CMS-like system to edit text, in the interest of better taking advantage of limited time and resources. Instead, we opted to use a collaborative text editor.

Since we wanted to make it easier to incorporate the edited text and automate as much of it as possible (to work within the constraints of the iterative process we were pursuing) we went for StackEdit, a lightweight, open-source Markdown editor that plugs directly into GitHub and enables huge flexibility. By plugging this into the main repository, we were able to keep both code and text edits harmoniously in the same file structure.

4 Validation

Validating the visualization sets involved two rounds of open-ended exploratory user tests and a final round of user experience rating along with blind tests.

4.1 First round

4.1.1 Test

For the first round of user tests, we only had the first visualization set in a state where it could be thoroughly tested, so we focused on it specifically.

5 subjects were given access to the website on a Mac laptop with little in the way of context, other than the draft title of the visualization and some words about Congreso Visible's mission. After leaving the subjects to use the site freely, we scored the users on subject matter comprehension and experiences using **yes** or **no** questions.

The results are as follows:

- The user understands the subject matter when asked to explain it (5/5)
- The user understands what each dot represents when asked (4/5)
- The user understands the beeswarm scale when asked to explain it (3/5)
- The user can identify opposition and government parties (5/5)
- The user attempted to interact with the chart (1/5)

We also gathered recurring comments from the users, ranked in the order of how many times they were received:

1) "The animation of the simulation is fun" (5) and "[and] eye catching" (4).
2) "The animation makes it easier to understand" (4).
3) "The scale on the last beeswarm plot is confusing" (4) "... at first" (3).
4) "The colors of the party can be confusing" (3).

4.1.2 Changes resulting from feedback

While the feedback was generally very positive, no user attempted to interact with the chart. This immediately made us take a different approach to developing the second set and simplifying the first one, towards a more "spoon-fed" approach where user interaction is not necessary but adds to the experience or helps deepen certain things.

The following changes were made to the first visualization set after this feedback:

- A color scale was added to aid the understanding of the scale in the beeswarm plot. We eventually extended this to the second visualization set too.
- Callouts were added to direct the user's attention.
- Some ideas relating to adding user interactions were discarded in favor of a simpler plot.
- A small helper text was added to encourage the user to interact with the chart.
- Conflicting colors were changed, out of which the parties with the most representatives kept their original color.

4.2 Second round

4.2.1 Test

For the second round of user tests, we focused on the second visualization set.

Once again, 5 subjects were given access to the website on a Mac laptop with little in the way of context, other than the draft title of the visualization and some words about Congreso Visible's mission. After leaving the subjects to use the site freely, we scored the users on subject matter comprehension and experiences using **yes** or **no** questions.

The results are as follows:

- The user understands the subject matter when asked to explain it (2/5)
- The user understands what each line represents when asked (4/5)
- The user understands the line chart scale when asked to explain it (1/5)
- The user can identify opposition and government parties (4/5)

We also gathered recurring comments from the users, ranked in the order of how many times they were received:

1) "There's too much going on" (5) or "there's too many lines in the same chart" (4) .
2) "I need it to be explained beforehand" (3).

4.2.2 Changes resulting from feedback

The feedback on this round was much less positive. This led us to rethink this chart in the spirit of less cognitive load

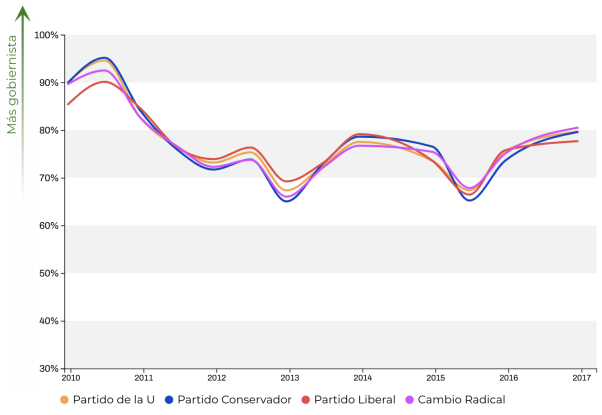


Fig. 3. Example of the reused color scale on the second set.

TABLE 1
Third round UX questionnaire results

The website is easy to understand.	5
The website is easy to use.	4.6
The first visualization set is easy to understand.	4.8
The second visualization set is easy to understand.	4
The website is fun to use.	4.6
The website contains too much text.	1.8

and overall just focusing on one data point at a time. This implied the following changes (among others).

- The same color scale for the first chart was reused for the second one.
- Each value line was added at a time instead of multiple at the same time, except when it specifically highlighted an insight to have them simultaneously.

4.3 Third round

The users were asked to use a special, redacted version of the site with no explanatory text. After a few minutes, they were asked to explain what each visualization set was showing. 5/5 participants correctly explained the overall gist of the visualizations. During this part, they were also scored in a similar fashion to the last test.

- The user understands the subject matter when asked to explain it (5/5)
- The user reaches the end of the site without direction (3/5)
- The user attempts to interact with the data using the mouse in the first half (3/5)
- The user attempts to interact with the data using the mouse in the second half (1/5)

Then, after being given access to the full, text complete site, the users were asked to answer more Likert scale experience questions. The results are outlined in Table 1.

The scale is as follows:

- Disagree
- Somewhat disagree
- Neutral
- Somewhat agree
- Agree

5 Conclusion

The trend towards more data collection and the increasing power of data insights means that some fields of knowledge that traditionally dealt with qualitative, multifaceted or descriptive-rather-than-measurable information, find themselves collecting unprecedented amounts of quantitative data. Fields like the political sciences or the social sciences as a whole find themselves in the position of being able to seize this potential of data and data visualization in a similar fashion to what the exact sciences have been able to do for centuries.

However, the potential of these incredible amounts of data tends to be untapped outside of the realms of what is possible within the profit margins of large corporations, since there is little incentive to spend large amounts of time and money investing in data analysis projects that are not conducive of larger revenue or profits. This means if we're to do this in a non-commercial environment, one either has to rely on open data or have to manage the collection oneself.

As is the case with other governments in developing countries, the Colombian government has been slow to implement open data into their policies. There has been some progress done with open resources from a handful of governmental or government-adjacent organizations, as well as initiatives by the Ministry of Information Technologies and Communications (MinTIC) that strive to publish open government data online. However, these often involve incomplete or otherwise hard to use data, is largely never updated or otherwise obsolete.

This means that projects like those undertaken by NYT Graphics and similar organizations are much harder in the context of Colombian politics and society. These depend on timely, up-to-date data being published by the government, which is hard to come by in Colombia, which means that the task of data collection has to fall on private shoulders if it is to be relied upon.

Congreso Visible, which encompasses a team of political scientists and journalists, until now lacked the ability and manpower to turn this data into interactive data visualizations. We believe we are among the first teams in this field dealing with Colombian political data in this format and we look forward to it being published and receiving feedback from the public.

Acknowledgments

The team would like to thank the team at Congreso Visible, specifically Felipe Botero, Beatriz Gil and Michelle Mora, whose help and prompt feedback was invaluable in this process, as well as Jos Tiberio Hernndez from IMAGINE Research group.

References

[1] M. Bostock, V. Ogievetsky and J. Heer, D3: Data-Driven Documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011.

[2] E. Mendes Rodrigues, N. Milic-Frayling, M. Smith, B. Shneiderman and D. Hansen, Group-In-a-Box Layout for Multi-faceted Analysis of Communities. *IEEE SocialCom GIB*, 2011.