A “crowdsourced” archive: Using Wikipedia to build a database of Indonesian cabinet members

Derrick Gozal

Master’s Candidate, 2023

Columbia University

Quantitative Methods in Social Science

Master’s Thesis

Abstract

While we often think of politics primarily in terms of formal institutions and the powers they confer, political actors themselves – independent of any office they may hold – as well as their interpersonal relationships with each other, also matter. Unfortunately, data on political elites – one of the empirical building blocks of political analysis of the latter sort – is often lacking, especially for developing countries. This study aims to help fill that gap by developing a method for building a database of political elites using scraping-based techniques to extract data from public sources of information. It does this by conducting a "test run" of this method by applying it to build a database of Indonesian political elites – particularly cabinet members – using data extracted from Wikipedia. The resulting database was quite broad – covering 90.2% of all the cabinet members that have ever served in Indonesia since its independence in 1945 – and rich in informational content despite a non-trivial degree of missingness (particularly "unknown" missing values). Additionally, the code and methodology used to generate it also looks to be relatively replicable, with some relatively straightforward manual cleaning involved. While the resulting database does not match the depth and "cleanliness" that datasets generated by expert-based methods can sometimes have, it does stand as a less cost and time-intensive alternative that nonetheless still has a significant degree of breadth and richness.

Table of Contents

1. Introduction

To many, it is perhaps second nature to think of politics primarily in terms of offices and officeholders. One becomes a minister and exerts influence as long as he/she retains office. The moment the office is separated from the individual however, we slowly but surely develop a kind of political amnesia directed specifically towards this individual. What this reflects is an assumption that offices and formal institutions are what confer all power on the individual, and not the other way around. "Kissinger was powerful primarily because he was Secretary of State, and not because of anything inherent to the man or his relationships," so goes this school of thought. Political reality of course, is rarely so simple. Elder statesmen continue to tug at the reins of power long after their official "retirement," oligarchs without any formal office exert their influence through money and guile. These phenomena are likely to be even more accentuated in political societies where informal power structures and institutions hold greater importance than their formal counterparts. A more nuanced understanding of politics then, would allow us to approach the study of political societies not solely through the lens of formal institutions, but from the perspective of these individual political actors themselves and how they interact with each other and these institutions as well.

One of the primary empirical building blocks of this more agent and network-based approach is information on these political elites: their biographical and demographic traits, their values and ideologies, their career trajectories, and their various formal and informal relationships. As will be explored further in the literature review however, this empirical data is often lacking and uneven, particularly for countries outside the developed West. Additionally, data collection for building up such databases can be tricky, especially in countries where information on such elites may not be available to the public. Methods that rely on expert interviews and some form of journalistic sleuthing are invaluable – especially in these countries – and have been used to construct such databases, but can often be quite costly and time-consuming.

The explosion of both computational techniques and information available on the internet however, provide unique resources that may help provide some solutions to the problems outlined above. Indeed, if we can leverage web scraping techniques in some form, it is more than possible to automate the extraction of large quantities of data from online sources. And if these online sources prove to be reliable sources of information on political elites, we can very well turn these techniques into a more time and cost-friendly alternative to the database construction methods outlined above. In essence, by relying on the Internet's very nature as an open and self-updating information bank, we open political research to the possibility of essentially "crowdsourcing" its data.

The focus of this study then, is on implementing a "test drive" of these techniques by attempting to create a database of Indonesian political elites using data extracted from Wikipedia. The choice of Indonesia is motivated by my own personal familiarity with the country; as well as the fact that it is also an exemplar of that family of countries outside the developed world where empirical data on its political elites remain relatively lacking, and where a preponderance of information political institutions makes the agent and network-based approach outlined above particularly suitable. I settled on Wikipedia – its Indonesian language version to be precise – as my main source of information due to its balance between accessibility and comprehensiveness, factors that I shall discuss in further detail in the segments below.

2. Literature Review

2.1. Grey eminences and bossy oligarchs: Defining political elites

In his famous 1962 paper, classicist Moses Finley laid out a series of astute observations on the nature of Ancient Athenian democracy, and in particular on the structural role of the demagogue. Frequently cast as a self-serving villain that riles up popular emotion for ill-advised ends, the demagogue occupies an inglorious position in most classical sources on Ancient Athens. Playing the part of the revisionist, Finley challenges this argument, attempting to rehabilitate the demagogue as a structurally crucial part of the Athenian political machinery. The demagogue, with his ability to alter Athens' political winds by his sole reliance on a glib tongue, represents that most crucial element of Ancient Athens' political robustness: political churn. In a world where any effective direct appeal to the masses had the ability to break any political faction's stranglehold on power, factions did not coalesce into permanent ruling classes but remained fluid and ephemeral. Thus did Athens escape the political stagnation that often followed factional dominance in most Greek cities of the time, a phenomenon that Finley called *stasis*.

Finley's commentary on Ancient Athens has an important lesson for modern political science: elites, and their relationships with each other, matter a great deal. If we are to define elites, as Mosca (1939) did, as a small group wielding disproportionate political power, this claim seems to hold intuitive sense. By Mosca's definition, we might say that Ancient Athens never really produced a long-lasting elite class because of its institutions, among which the much reviled demagogue was a key feature. Conversely, one might also see in the demagogue's disproportionate power to move political winds with sheer oratory the seeds of something resembling a political elite, albeit perhaps a non-durable one. We might even imagine contemporary counterparts to these demagogues of old: perhaps podcasters who politicians compete to have a broadcasted conversation with, respected religious leaders who swing votes with declarations of support or condemnation. These thought experiments should make it clear that the actions, or at least the theoretical place, of elites may have great bearing within political systems, and are thus worthy of study.

Defining the term “political elite” itself however, turns out to be a fairly tricky proposition. While it may seem intuitively easy to grasp that there is a disproportionately small ruling class in political societies, things get muddy the moment we try to get more granular with our definitions, something that will have practical implications the moment we have to make judgement calls about who is and isn’t a political elite. Bottomore’s (1964) definition of an elite class as one which exercises disproportionate power appears to be the most in line with what we think of when hear of the term. However, this forces us to ask a follow-up question: what constitutes power? Is it wealth, military dominance or perhaps even something less tangible like charisma? This gets trickier if, as Dahl (1958) critiques, power is conceptualized as being so omnipotent that it effectively becomes invisible. In this scenario, the truly “powerful” political players are those unidentifiable forces lurking in the shadows. Others like Mills (1959) define elites in terms of the formal institutional positions that they hold. But this, as Zuckerman (1977) rightly points out, lacks cross-national comparability. If institutions change with different national systems (and they certainly do), then we are left without a uniformly comparable definition of elites. This also raises another issue: how does this definition apply in systems where informal institutions dominate. Still others, such as Dahl (1958) use more concrete criteria such as cohesion. Taken to the extreme, this means that a group of politically influential but perpetually bickering aristocrats cannot be classified as elites.

How do we escape this morass? Zuckerman (1977) proposes a solution, based on the writings of both Mosca and Pareto. He begins, as Mosca and Pareto do, by accepting the basic intuitive idea of a political elite, reasoning that “most (individuals) are not involved in political life.” He then suggests that any conception of elites can only make sense within the context of a broader theoretical system in which elites play the role of one of the chief monopolists of power. What this looks like exactly depends on the specific theoretical system that one elaborates. As such, defining elites is only sensible vis a vis the structure of the theoretical system that we choose to situate them in. “First comes the system, then the elite,” so to speak. We will use this definition of elites as a conceptual starting point, and try to take things from there.

With the above in mind, we may start by trying to sketch out a basic theoretical structure of differing political systems and then try to conceptualize how we might define elites in each case. One way of juxtaposing these systems that one may occasionally encounter in the literature is to classify networks and institutions as two different kinds of social/political arrangements (Beteille, 2009). It is certainly true that this is by no means a sharp delineation. Finley’s own example above is an excellent case of a blurring of the lines. The institution of Athenian democracy shapes how elite networks form and operate, but it is the character of the demagogue that helps maintain the institution as well. However, it is also not unreasonable to posit that different political systems may emphasize one of these two forms of arrangement in varying degrees. In other words, just as certain political systems may be defined primarily by their formal institutions: their parliaments and cabinets; we may also imagine systems where power-sharing agreements and patronage networks have a more decisive influence than formal institutions.

One might see how we could define elites differently under the pure forms of each political system. Under the former, one might analyze elites in terms of their formal rank and the formal institutions that they occupy. We may analyze their behavior in terms of say, the formal powers vested in the specific positions they hold. What emerges is a view of elite behavior that is subsumed by the broader institutional structure they occupy, that is based on the formal rules of the institutional game that makes up the political system we are studying. We see a very different conception in the latter world of "networked" systems. Here the formal rules that dominated the institutional world are somewhat de-emphasized. Instead of the formal powers of the position they occupy, we may analyze elites through their direct connections with each other instead. Power here lies not in amassing formal positions, but through monopolizing connections with critical individuals and networks. Again, systems that are purely institutional or network-based are merely pure forms that rarely (if ever) exist in the real world. Most systems lie on a spectrum that emphasizes one form or the other.

This juxtaposition however, is complicated by an existing imbalance between the empirical evidence available for each kind of system, which has ramifications for studies on elite politics. Indeed, while studies on institutions are abundant, there remains a noticeable gap between these and hard data on political individuals, which comprise one of the empirical building blocks of analyzing politics from a less institutional and more network and individual-based approach (Gerring et al., 2014). Fortunately, there has been an increase in these kinds of elite databases over the past few years. Examples include Gerring et al's (2014) own Global Leadership Project, which provides a comprehensive cross-national database; and several in-depth databases on the Chinese Communist Party elite from Shih, Meyer and Lee (2016) as well as Jiang (2018). Other good examples include the EurElite (Best & Edinger, 2005) and SEDEPE (Dowding & Dumont, 2009), which focus on several Western democracies. However, there remains a problem with regards to developing countries, for which databases such as these remain relatively rarer. In the cases where some sort of individual-level data on political elites exists, these are often limited to heads of states and cabinet ministers. Data on broader swathes of the elite (such as perhaps legislators, members of the military and police apparatus, etc.) is often missing.

2.2. The case of Indonesia

This lack of data can be particularly obfuscating for countries such as Indonesia, where patronage networks appear to have the upper hand over formal institutions in determining politics. Indeed, Aspinall et al. (2018) note that Indonesian parties function more like cartels founded on pork barrel agreements rather than actual ideological positions. Already this diminishes the classical institutional role of parties as formal vehicles for popular ideologies. This dynamic is also particularly salient at a local and regional level. Various studies, such as those by Rusnaedy and Purwaningsih (2018), Burchanuddin et al. (2021), Muksin, Purwaningsih and Nurmandi (2019) and Ardiman (2022), paint a portrait of a regional electoral landscape characterized by weak parties and strong dynastic elites. Party and ideological discipline are often subordinated to the economic and social capital of local elites. Chalik and Latif (2020) even describe scenarios where parties who are ideological opponents will occasionally band together at the regional level for reasons of pragmatism.

2.3. Surveying the landscape of political elite databases

Taken as a whole, these works make a strong argument for taking a more agent and network-oriented view – focusing on individual elite relations instead of formal institutions – when analyzing the Indonesian political system. It also makes the task of creating a database of Indonesian political elites particularly important. To be clear, certain sources already exist. The

WhoGov dataset by Nyrup and Bramwell (2020) contains a historical database of cabinet members from various countries (Indonesia included). However, this still falls prey to Gerring et al's (2014) criticism of having an overly limited definition of elites, since this database only includes cabinet members. Gerring et al's (2014) own Global Leadership Project database is much more comprehensive and contains a wide variety of political elites. In fact, one of its biggest advantages is that it also tries to include informal holders of power in its database of elites. However, this database only appears to include information on active politicians. This makes it less useful for historical analysis. Additionally, neither of the aforementioned datasets include information on potential network ties such as familial relations or specific educational and organizational information (which can be used to triangulate cliques based off common organizational memberships or alma maters).

Collectively, these factors indicate that there remains plenty of room to broaden our existing constellation of empirical data on Indonesia's political elite. Gerring et al (2014) relied primarily on questionnaires filled in by country experts. As mentioned above in the introduction, this approach poses resource and time constraints that our computational "crowdsourced" method hopes to circumvent to some degree. Some comparative questions between this computational method and Gerring et al's interview-based approach need to be kept in mind however. For instance, do the time and resource savings that computational techniques allow for come at the cost of some of the depth and nuance that expert surveys might provide? We may see some hints in the WhoGov dataset by Nyrup and Bramwell (2020), which seems to rely on computational techniques as well. In this case, OCR is utilized to obtain a digital version of the *Chiefs of State and Cabinet Members of the Foreign Governments* publication by the CIA. The extracted raw data appears to be relatively clean, and is then further refined using matching methods and machine learning. Additionally, the researchers still needed to have recourse to manual methods of cleaning and verification at the end to be doubly sure of the data's veracity. There are several main takeaways here. First, that the original source from which I am scraping information from will have big ramifications for how I will have to design my extraction methods. Wikipedia will have its own idiosyncrasies distinct from the collection of texts that Nyrup and Bramwell (2020) used. Secondly, in spite of all the computational methods introduced for scraping and cleaning the data, it might still be necessary to resort to some manual cleaning at the end for final verifications. Wikipedia's own lack of regular structure may make the need for this more pressing.

With that in mind, we turn to Wikipedia's own reliability as a source of information. There has been some literature written on the virtues and vices of using big data – particularly those of the crowdsourced variety – in the social sciences, and we can use the intellectual framework they provide to guide our evaluation of the computational methods we will be using in this study. Porter, Verdery, and Gaddis (2020) write of the three "V's" that social scientists generally look for: volume, value, and validity. While big data delivers well on volume, how it performs on value and validity is more uneven. Indeed, these are features inherent to the "organic" nature of most big data, which are not generated with the explicit purpose of answering any specific research question. This "organic-ness" allows for the constant accumulation of data by the routine operations of everyday society, which explains volume. Without the guiding hand of a researcher to curate this data generating process however, there is always the risk that big data fails to meet the value and validity checks necessary for the specific research goals the social scientist wishes to pursue. Discussions of Wikipedia as a source of information will be included in section 3.2 below, and will be done keeping these considerations of the three V’s in mind. Additionally, Porter, Verdery and Gaddis also mention the importance of proper documentation and reporting when it comes to using big data, particularly in light of the above concerns on value and validity. Gerring et al’s (2014) attempt to record and be transparent about the estimated degree of completeness of the database is an example of this, and I will be including similar documentation of my work in line with these concerns.

3. Data and Methods

3.1. Which political elites?

The first order of business is to decide precisely what kinds of political elites I will be focusing on in this study. Following Gerring et al’s (2014) notes above, we can define political elites based on their position in the government’s various formal branches: executive, legislative, and judicial. Other formal governmental staff such as ambassadors might be included in this list. Additionally, members from other institutions and political organs such as parties, the military and police might play an important part in elite politics and should be included as well. Leaders of civil organizations with disproportionately heavy political influence such as the Islamic mass organizations *Nadhlatul Ulama* and *Muhammadiyah* may also be included in this list. Lastly and most tricky; informal political elites with no well-defined positions such as informal advisors, *eminences grises* lurking in the shadows, or business tycoons with political ties; may all be classified as elites as well.

The availability of information for each of these categories of elites differs greatly, and will be an important consideration going forward. Executive-level elites tend to have the most complete information on them available. Leaders of important institutions and organizations such as parties, the military or influential civil organizations should also have a reasonable amount of information available. While rosters of the names of legislative elites should be readily available, more detailed information on each of these elites may be much more uneven. The most challenging type of elites to gather data on however, are by far those who hold no formal positions. In patronage-based systems such as Indonesia, these “informal elites” may play a disproportionate role in the political system, which makes this informational scarcity a rather serious issue that should be considered. Keeping this in mind, we will first try to create our database focusing solely on executive-branch elites. The primary goal here is to first test our scraping techniques on cabinet members, and then to see if these can be expanded to include other elites. After all, going beyond the cabinet level is one of the primary motivations for this study.

We turn now to the next question: what sort of information on each elite should we have in our dataset? At the most basic level, the dataset should include basic biographical information on each elite. This should include their dates of birth and death, gender, and religion. Since the Indonesian government doesn’t officially record one’s ethnicity, a better option might be transcribing each elite’s place of birth. Not only does this allow us to make a rough guess of one’s ethnicity, but also to gauge out potential geographical cliques and relationships. Beyond that, relevant political information such as political party affiliation and a history of the positions the elite has held should also be included. Educational history would also be included, since they might reveal cliques centered around certain educational institutes or even how certain educational institutions may act as political pipelines. Finally, the database will also try to incorporate each elite’s family history to the greatest possible extent. This will allow us to trace familial relationships between elites in the manner outlined in the literature review above.

3.2. Why Wikipedia?

As mentioned earlier in the literature review section, the question of what source to use for the database is an important one, and one that we will now turn to. Again, keeping in mind the discussion in the literature review on what are missing from the GLP and WhoGov datasets, the most important features of our data source should be that it has some degree of historical comprehensiveness (with data on elites stretching back to at least Indonesia's independence in 1945), and that it contains information on potential network ties such as familial relationships, as well as educational and organizational memberships. An interesting option that might have all of this is news sources. Indeed, the local media often conducts spotlights on newly appointed ministers, which occasionally includes their career and educational history, as well as their family members. Several blogs run by journalists such as Yosef Ardi are also rich with information on the labyrinthine networks and connections that tie the Indonesian elite together. Combining some form of scraping and a combination of named entity recognition (NER) and relation extraction (RE) on Indonesian news sources may work well to extract information from such sources. These are essentially machine learning-based NLP models that identify entities and draw relations between them in texts. There are two primary roadblocks here however. Firstly, there exist few NLP packages that are designed to conduct the above-mentioned analysis on text written in Indonesian. While it is possible to translate these texts into English and then use English-based NLP tools, the spotty quality of translations means that our NER and RE models may not have the best accuracy. Additionally, several of these sources, such as the aforementioned blogs, are occasionally locked behind paywalls that limit access. Most also do not have APIs that might allow for easier access.

In light of these limitations, an alternative source that one might consider is the Indonesian language version of Wikipedia, which is much more comprehensive than the English language version when it comes to Indonesian political elites. Indeed, the Indonesian version of Wikipedia is both historically more comprehensive – with information on various elites stretching back to the 1940s and even earlier – and also contains many of the fields missing from the other databases mentioned above, such as family relationships and career history. A rough accounting shows that around 90.7% of all Indonesian ministers that have served on a cabinet from the nation's founding in 1945 have Wikipedia entries. When we limit this to only ministers who have served in cabinets from Suharto's New Order period onwards (post-1965), this number shoots up to 99.4%. In terms of volume then, Wikipedia seems to perform quite well.

What of validity and value? Firstly, while Wikipedia remains broadly reliable, one must still keep in mind that this is a publicly-maintained database rather than an "official" dataset. The issue of how reliable the information in Wikipedia is remains an open question. Fortunately, most of the information I seek to extract consists of simple, public biographical facts that are easy to fact-check such as date of birth and political party affiliation. As such, we may assume a reasonable degree of validity.

Additionally, we cannot assume completeness from this data. For example, say we want to extract information on politician A, it is possible that all the information we are extracting from his Wikipedia page is by no means complete. Our database might list all his positions in the legislative and executive branches but omit his previous experience as a diplomat in various countries. It is with regards to this issue of missingness that our Wikipedia-based dataset will be most deficient vis a vis validity and value, particularly when we want to conduct analysis that assumes completeness of data. To return to the example of politician A: say we want to look at the career trajectories of politicians to see if there is some sort of *cursus honorum* in the Indonesian political order. It might be the case that starting out as a diplomat is often a stepping stone towards a higher ministerial-level position, or perhaps even a "reward" after serving in a ministerial position. Such a phenomenon however, will not be visible in our dataset if politician A's occupational history is incomplete in his Wikipedia page.

In spite of these two limitations however, sourcing data from Wikipedia strikes a balance between comprehensiveness and feasibility, which is what eventually pushed me to select it as this study's main data source. Since Wikipedia has an API, I will be using that to extract all of the data from it. Additionally, while it may have been ideal to use a combination of NER and RE to extract all the relevant information from the main body of each Wikipedia page, as mentioned previously, many NLP functions do not have packages designed for the Indonesian language, making the use of these models somewhat challenging. As such, I decided to rely on the infoboxes that populate every Wikipedia page instead. Here information is semi-structured, with each piece of information labelled and categorized, albeit not in the most uniform and consistent way. For the most part, I used regular expressions to clean the data extracted from these infoboxes. I also developed an iterative function that looks up each individual's family members and tries to add them to the database if they are not already included. This function repeats and tries to do the same to these newly added family members until it exhausts all available relations. This allows for the creation of more comprehensive family trees and long chains of relations. On a final note, while I initially intended to automate as much of the extraction and cleaning process as possible, Wikipedia's lack of a strict page template means that the way information is stored often varies unpredictably. This necessitates a degree of manual cleaning, as will be further elaborated in the segments below.

3.3. Creating the database: A step-by-step guide

The following is a detailed, step-by-step accounting of how the database was created:

1. The first order of business here was to come up with lists of elites to look up. I started by creating a table listing all the cabinets that have served since Indonesia's independence in 1945. Since Wikipedia has pages dedicated to listing members of all of these cabinets, I simply used the list of cabinet names to access all of the pages that list the members of every cabinet. A screenshot of one such page of lists is provided in **Appendix A.**
2. I then extracted all the names with hyperlinks in every one of these cabinet list pages, with the idea being that entities with available Wikipedia pages would have accessible hyperlinks. The catch here though, is that this would also capture non-person entities with hyperlinks such as "Indonesia" or "The Natsir Cabinet." To account for this, I implemented a filter that drops any entity which does not have a birth date field in their Wikipedia page, with the assumption that only persons have a birth date field. Also, since some individuals serve in more than one cabinet, I also built in a check that does not add individuals that have already been extracted previously. Finally, I also needed a unique identifier for every entity (not only individual persons, but other entities such as political offices or educational institutions as well) since they may be referred to by different names in different instances. For example, Sukarno may be referred to as Soekarno in various instances, and the Minister of Defense may also be referred to as the Minister of Defense of Indonesia. Each Wikipedia entity with its own unique page has its own unique Wikidata ID, which I decided to use as the primary unique identifier to link entities across my database.
3. Next, I extracted information from the infobox segment of each person's Wikipedia page. Parsing through and cleaning the information available here was by far the most time-

consuming portion of this study. Again, I mostly used regular expressions to clean most of the extracted data here. The specifics of this process are convoluted and varied with the different structures of each field. Further details on this can be viewed in the separately provided **Code Appendix**.

4. However, since Wikipedia's structure is not very consistent (examples of some of the raw text extracted from Wikipedia is provided in **Appendix B**), there was always the risk that my parsing and cleaning code would not capture some of the data with more idiosyncratic structures. As such, I built-in a check function that would mark "suspicious" entries. For instance, I would mark any supposed person entry when the Wikipedia page it directs to does not have a birth date, suggesting that this entry may not actually be a person. The reason for doing this is that manually checking the entire database row-by-row can be incredibly time-consuming. Having these checks allows us to cut down on manual cleaning time by instantly zoning in on the most suspicious entries only.
5. Once I've added all the relevant information on each person and checked all of them, I tried to implement an iterative function that tries to expand the database by also adding entries for their family members. In essence, this function tries to look up whether any of our original list of politicians' family members already exist in the database. If not, it looks up their Wikipedia entries and repeats steps 3 and 4 above to populate the database with their information. This function then repeats this with the family members of those newly added persons, and continues until it exhausts all the available family relations on Wikipedia. This iterative function resulted in quite a significant expansion of our dataset, where an initial collection of 685 unique persons expanded to around 1048 once this function was implemented.

6. At this point of the process all of the automated extraction has been completed. What remains to be done is some final manual checking, as well as the translation of these datasets into English. As mentioned above, the first was done by filtering for rows that were marked by the aforementioned checking function and by manually validating them, altering any if any mistakes were detected. Translation was achieved by using alias tables, where I extracted the unique entries for everything that needed to be translated: mostly educational institutions, parties and offices/positions. I would then create a new column containing an English alias of each entity's name and join this with our original table. Creating this alias table also gave me the opportunity to add further custom details to each entity that could help enrich analysis. For instance, I added labels to each position specifying their nature (eg. Executive, Legislative, Military etc.).

The above is a general rundown of how I went about constructing my database. The full code with all of the nitty-gritty will also be provided in a separate **Code Appendix**.

4. Results

4.1. Overview of database

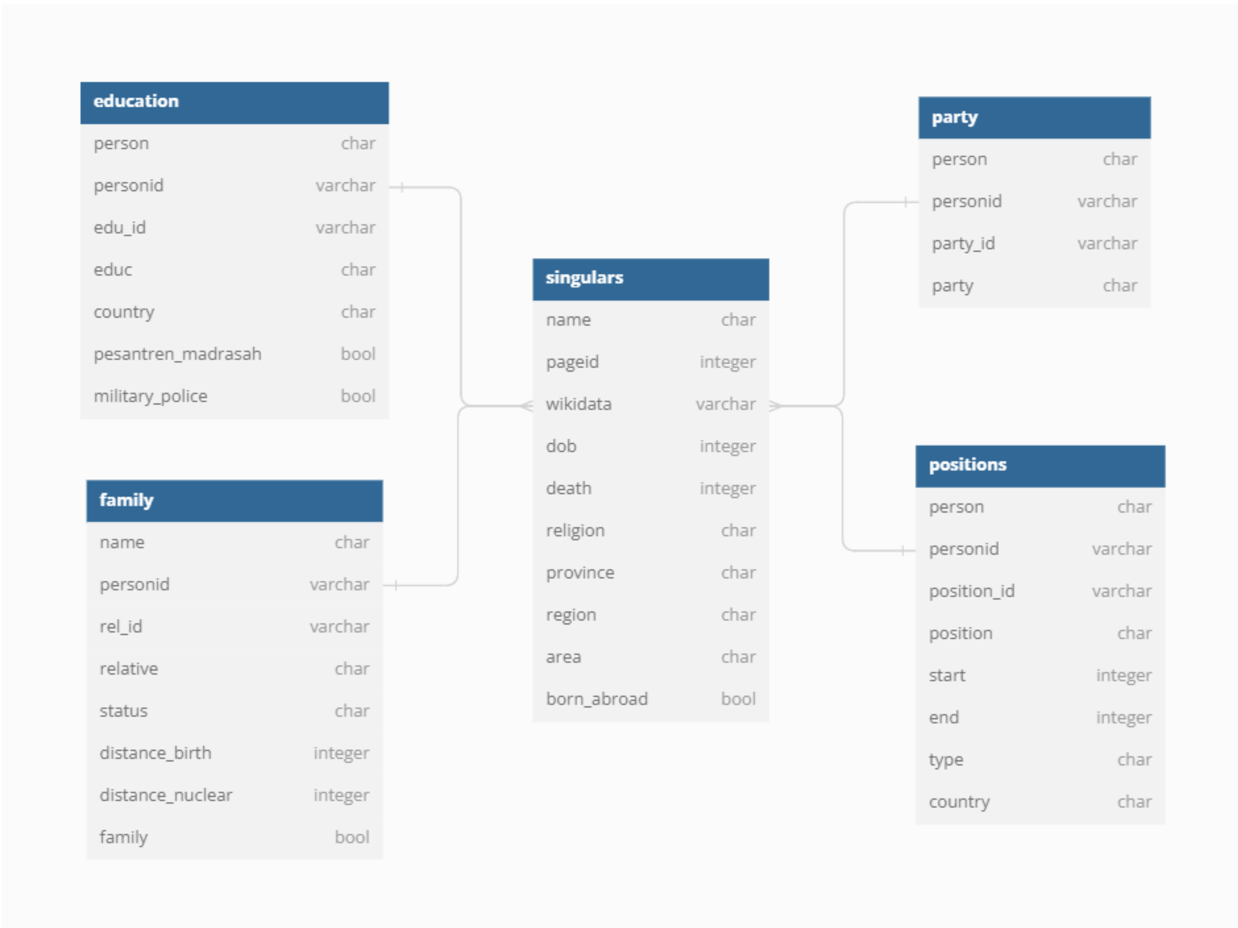
For the final output, the full database consists of five different tables. I've created a star schema with a primary table containing information on all our individual elites linked to various other tables containing other, "multi-row" attributes for each individual:

- **Singulars table:** This is the primary table containing basic, "single-row" attributes of each individual. This includes attributes such as each elite's name, date of birth, date of death, birth province, and religion. These are "single-row" attributes in the sense that each individual elite can only have one name, date of birth, or religion. In the final version of this table, I also added a dummy variable specifying whether a particular individual was born outside of Indonesia or not.

- **Positions table:** This table lists all the official positions held by each individual, as well as the start and end dates of holding each position. As a single individual can hold more than one position, this is a "multi-row" rather than a "single-row" table. Additionally, I also added a column detailing the nature of each specific position (eg. Executive, Legislative, NGO, Military, etc.). Also, since the above iterative function managed to catch several Japanese politicians who are somehow related to Indonesian elites, I also added a column specifying the country each position is based in.
- **Parties table:** Lists all the parties that each individual has been a member of. Again, this is a "multi-row" table. Unfortunately, this does not include the dates of membership for each party membership.
- **Education table:** Lists all the educational institutions that each elite has been to. This is also a "multi-row" table. Again unfortunately, dates of attendance are not included. Certain details of each institution such as the country of its location are also included.
- **Family table:** Lists all the family members of each elite. This is also "multi-row" since each individual can have multiple family members. The recorded "distances" of each relation varies significantly. For instance, for a certain individual only nuclear family relations (eg. parents, spouse, children) may be recorded; but for another, more distant relations such as great-grandparents or even distant ancestors could be included as well. As such, certain fields specifying this relational distance were also included to help users better navigate this.

The following star schema illustrates the structure of the tables and how they relate to each other:

Figure 1. Star schema of all five tables that make up the final database



Additionally, a full metadata explaining each table can be found in **Appendix C.** Screenshots of each table can also be found in **Appendices D-H**.

4.2. Sample Analysis I: Geographical origins of cabinet members

This section and the following contain examples of the kinds of analysis that can be done using the generated dataset. In this section, I try to take a historical view of the geographical composition of cabinet members' birthplaces. Again, while data on ethnic identity may be hard to collect for individual politicians, we can use one's birthplace as a proxy for his/her ethnic identity. By observing the geographical composition of cabinet members' origins over time, we may be able to observe the emergence and decline of various geographical and ethnic cliques in the executive branch of the Indonesian government. We can start by approaching this from a very broad level:

Figure 2. Composition of cabinets over time, by area of birth

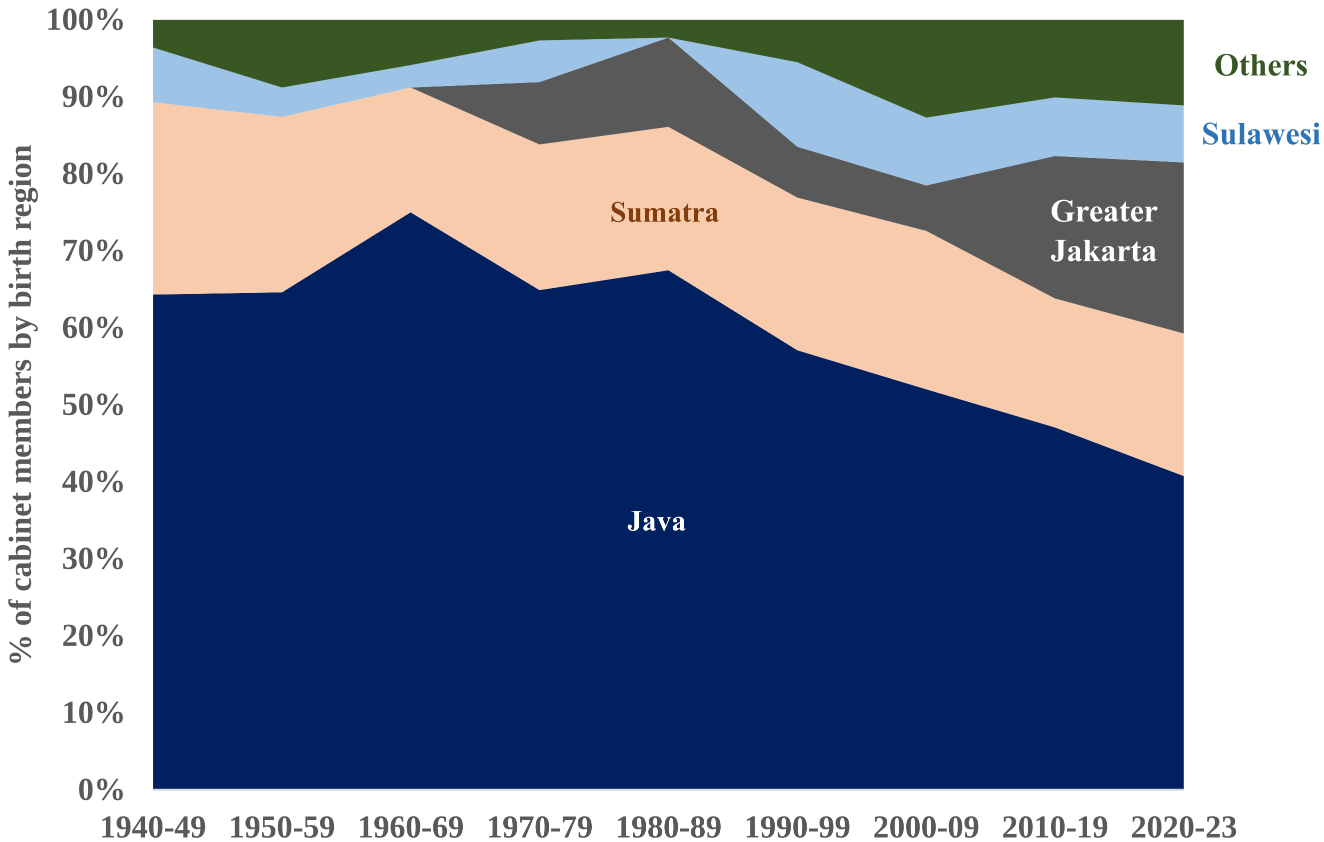
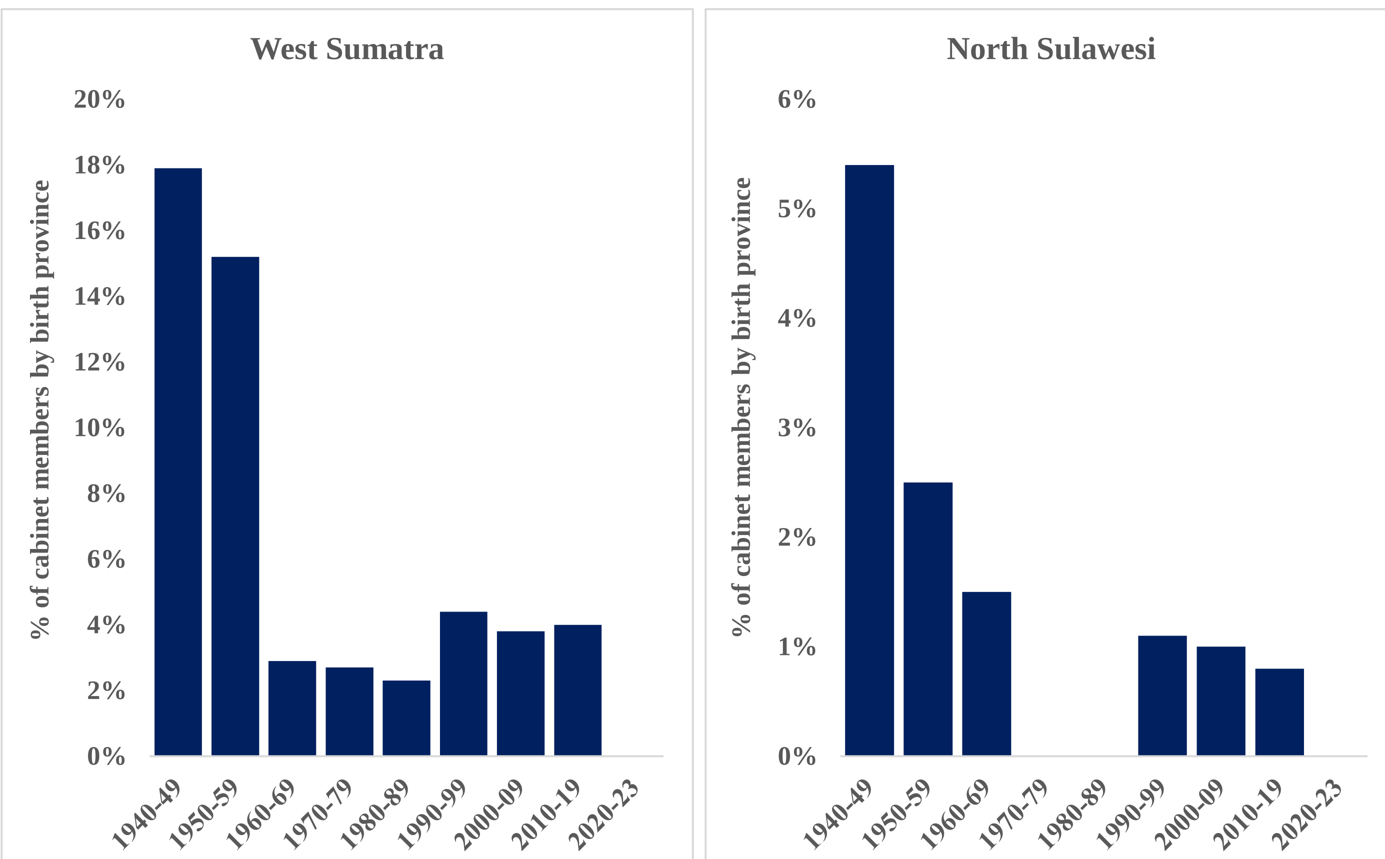
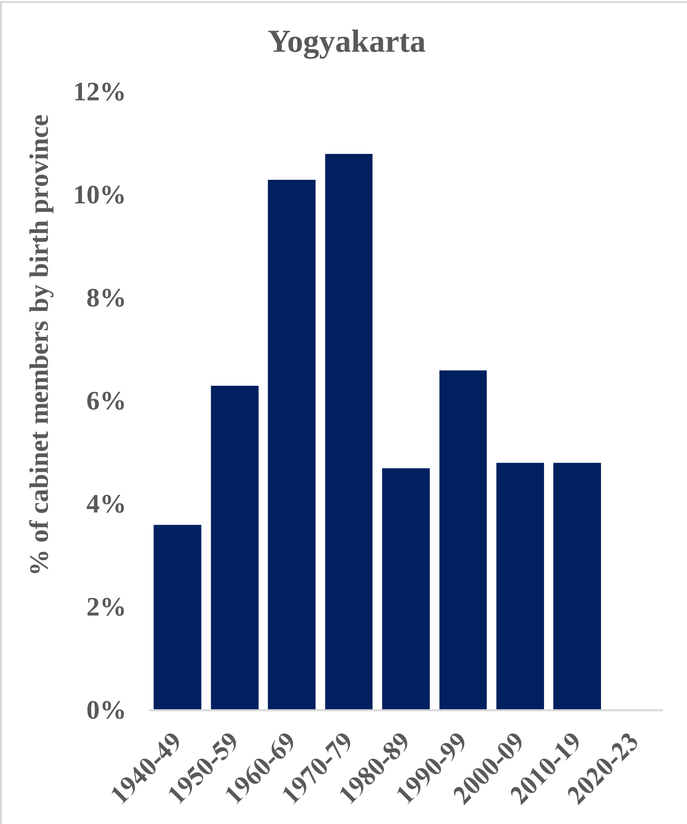
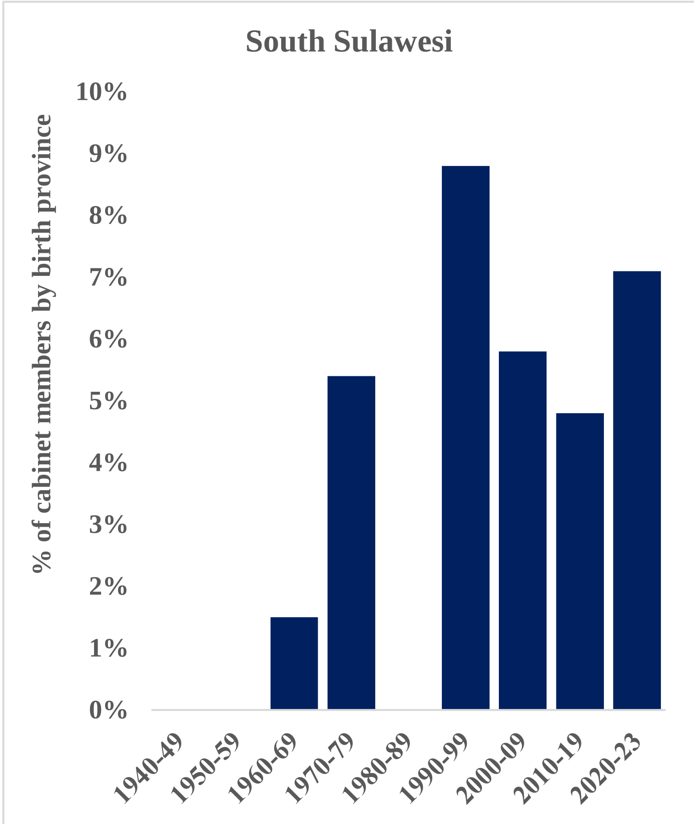


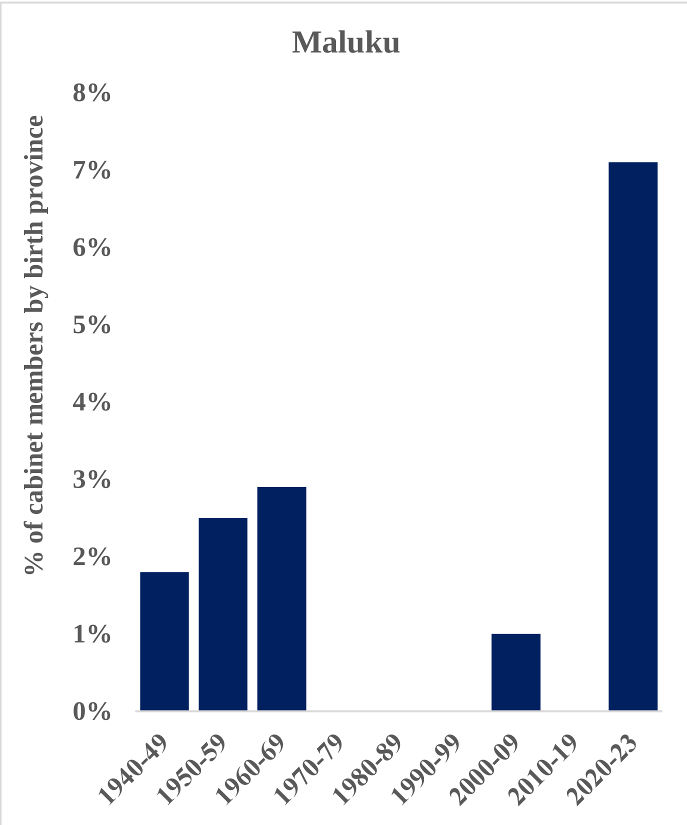
Figure 2 shows interesting trends with regards to the geographical composition of Indonesia's ministerial elite. The most striking observation is the sharp decline of Java-born politicians in Indonesia's cabinets over time. As the country's largest population center, Java has long been the center of Indonesian politics. Does the above signal some form of decline in Javanese dominance over Indonesian politics? Interestingly however, the decline in Java's representation in the cabinet has also been flanked by a sharp increase in Greater Jakarta's representation in it. It may very well be the case that large swathes of Indonesia's old regional elites gradually transformed into Jakarta-based elites once they managed to secure important political posts in the capital. Case in point is Indonesia's first president Sukarno. While he himself was born in Surabaya, East Java, all of his children were born in Jakarta. It is likely that what we are seeing may not be the decline of Javanese representation perhaps, but to some extent a significant transformation of an originally Javanese elite into a Jakarta elite. The increased representation of Sulawesi and other regions in the cabinet alongside Java's decline however, does indicate that more diverse regional representation is also taking place to some degree. While this simple analysis does not provide any definitive answers to these questions, it does highlight the analytical ways in which this database can be used. For further illustration, a similar analysis but for specific provinces also reveals interesting results:

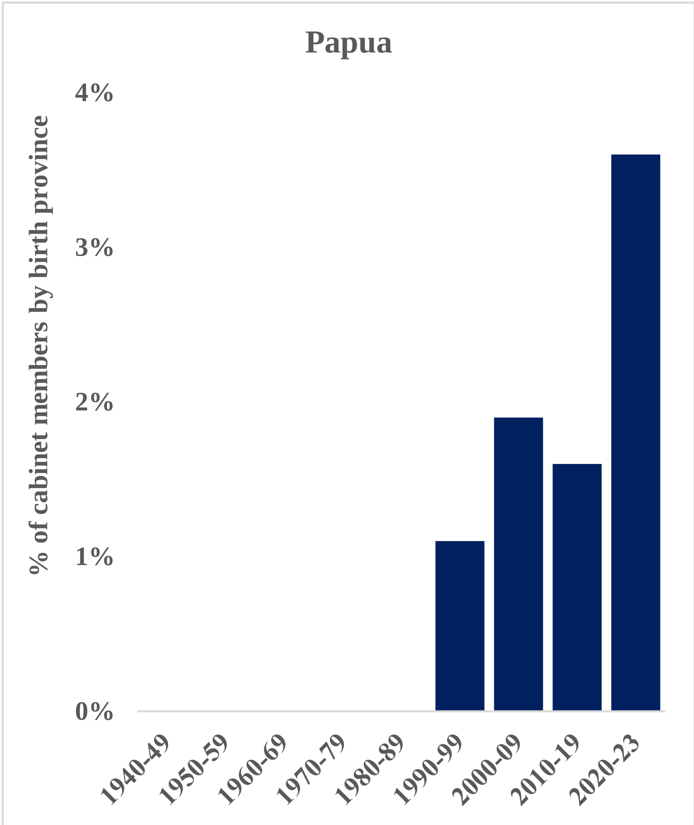
Figure 3. Composition of cabinets over time, by selected provinces of birth




Yogyakarta
% of cabinet members by birth province
12%
10%
8%
6%
4%
2%
0%
1940-49
1950-59
1960-69
1970-79
1980-89
1990-99
2000-09
2010-19
2020-23


South Sulawesi
% of cabinet members by birth province
10%
9%
8%
7%
6%
5%
4%
3%
2%
1%
0%
1940-49
1950-59
1960-69
1970-79
1980-89
1990-99
2000-09
2010-19
2020-23


Maluku
% of cabinet members by birth province
8%
7%
6%
5%
4%
3%
2%
1%
0%
1940-49
1950-59
1960-69
1970-79
1980-89
1990-99
2000-09
2010-19
2020-23


Papua
% of cabinet members by birth province
4%
3%
2%
1%
0%
1940-49
1950-59
1960-69
1970-79
1980-89
1990-99
2000-09
2010-19
2020-23

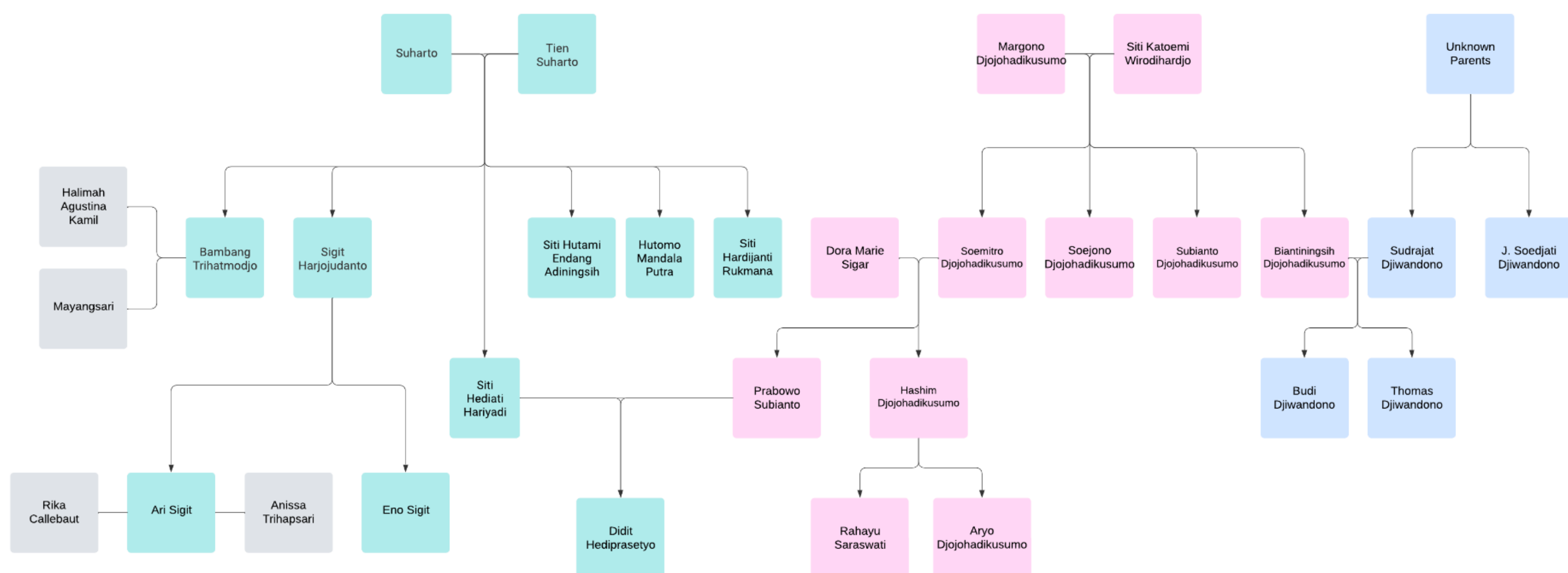
As we can see in figure 3, certain locales that appear to have wielded disproportionately large influence within the cabinet during the beginning of independence seem to have lost much of this influence over time. We see this to a certain extent with North Sulawesi, composed mostly of Minahasans. It may be the case that Minahasans, an ethnic group overrepresented in Dutch institutions such as the Royal Netherlands East Indies Army (KNIL), found themselves in greater positions of power during the early days of independence and gradually declined in representation as Dutch institutions and rule disintegrated in Indonesia. More striking however, is the observed decline in the representation of predominantly Minangkabau West Sumatra. This narrative of political decline features in certain popular imaginations of the Minang about themselves, where contemporary Minang influence is seen as but a ghost of what it was once was in an era once dominated by Minang giants such as Sutan Sjahrir, Tan Malaka, and Muhammad Hatta, all of whom could be considered pivotal founding fathers in one form or another (Razak, 2020).

Other interesting observations include the surge in Yogyakartan representation in the 1960s and 70s, which marked the beginning of the Suharto regime. It is probable that Suharto, being from Yogyakarta himself, had an inner circle that was dominated by other native Yogyakartans, at least during the early years of his rule. We also see increased representation of groups from Eastern Indonesia as we move into contemporary Indonesian history with the waning years of Suharto's New Order and especially in the current Jokowi cabinet; with groups from South Sulawesi, Maluku and Papua gaining considerable ground in cabinet representation.

4.3. Sample Analysis II: Constructing family trees

Based on the family table I've created, we can also attempt to reconstruct family trees. To illustrate, I've sketched out the Suharto family (Cendana) family tree below:

Figure 4. The Cendana family tree. We see here the intermixing between the main Suharto family (in teal) itself as well as two different families that have been grafted on to it by marital relations: the Djojohadikusumos (in pink) as well as the Djiwandonos (in light blue)



This family tree exercise is particularly instructive because it highlights both the strengths and weaknesses of our dataset quite well. On the positive side, I was able to construct much of this family tree solely using the relationships found in the family table. Some gaps do exist however. In particular, our table was recording some kind of relationship between the Djojohadikusumos and the Djiwandonos, with it recording Hashim Djojohadikusumo and Prabowo Subianto as uncles to Budi and Thomas Djiwandonos. However, it could not detect an unbroken chain of direct links between the Djiwandonos and the Djojohadikusumos. I had to manually look up information online confirming that Budi and Thomas' father was married to Biantiningsih Djojohadikusmo, one of Prabowo and Hashim's siblings, in order to complete this chain of direct links. The issue here is that Biantiningsih never made it into the dataset, most likely because she had no Wikipedia article. While this is an issue for constructing family trees composed of direct relations like the above, this should be less of a problem when conducting network analysis, where one can simply adjust the values on relationship ties to mirror their "distance," thus allowing non-direct familial relations (eg, uncle, great-grandparent) to be used when we have a direct relational gap like the above.

Despite this limitation however, we can see how analyses of this sort can be of use. Going through the entries of these individuals in the positions table for instance, will reveal that most living members of the Suharto dynasty remain quite active in politics. Hutomo Mandala Putra is now the chairperson of his own Berkarya Party. Prabowo Subianto is of course, one of the candidates for the upcoming presidential election. Hashim's own children Rahayu and Aryo both served as legislative councilpersons at some point in time, and the two younger Djiwandonos are currently serving as either legislative members or in leadership roles in Prabowo's Gerindra Party. Any suspicion that the influence of the Suharto clan has vanished in

Indonesian politics should be dispelled by all this, although it does appear to be the case that much of the political energy described above seems to gravitate more closely around Prabowo rather than Suharto's own offspring. While not quite dead yet, it does appear that the Suhartoist flame has passed from the Suhartos themselves to the cadet branches of the family.

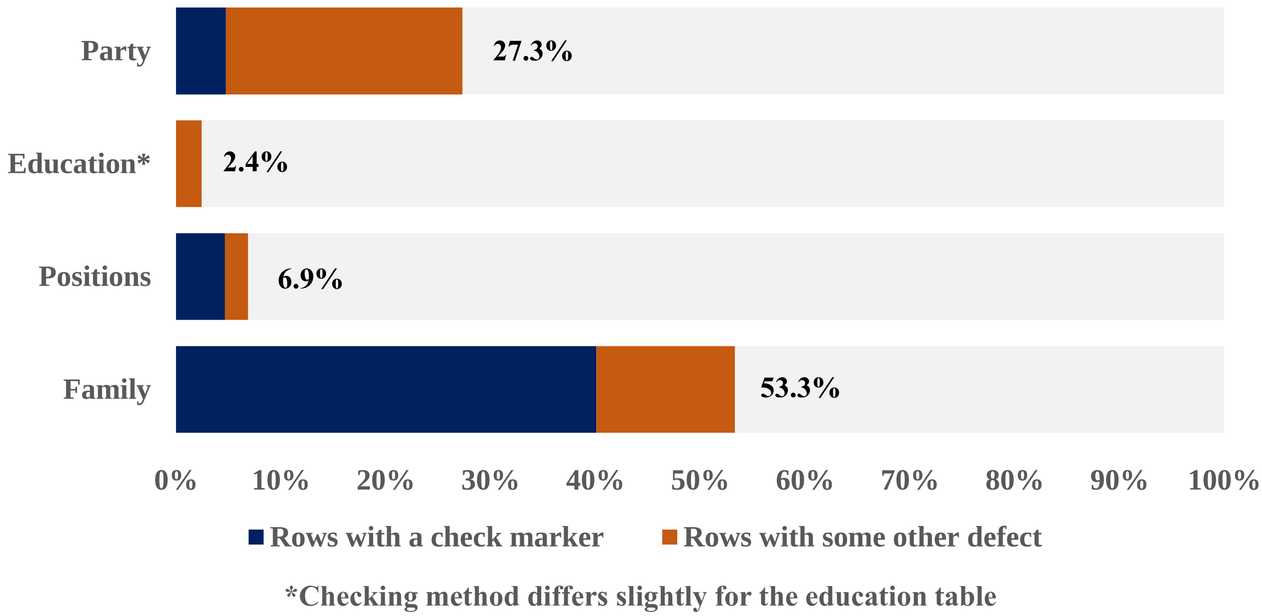
5. Discussion

In this section, I try to evaluate the generated dataset along several different dimensions, namely missingness and the degree of manual error checking required. Additionally, I will also try to identify which of the bottlenecks and limitations that show up in the generated database are those that will be largely surmountable with the right methodological refinements, and which reflect inherent structural weaknesses of the scraping-based methods used here.

5.1. Error checking: How much manual cleaning is necessary?

In section 3.3 above, I described the implementation of a checking function that marks entries that appear "suspicious" in our database, entries that we can then validate manually. Again, this is necessary due to Wikipedia not imposing any standardized template for its pages, which means that there is always a risk that any sort of parsing code relying on regular expressions might miss several unexpected formatting schemes. In any case, we can count the proportion of each data table that is marked by this function as a proxy of how much manual cleaning is necessary for each. In the course of cleaning, I also identified other features that clearly marked a row as suspicious and necessitated manual validation, features that varied with each different table. Figure 5 below shows the degree of manual validation required as indicated by these two different markers:

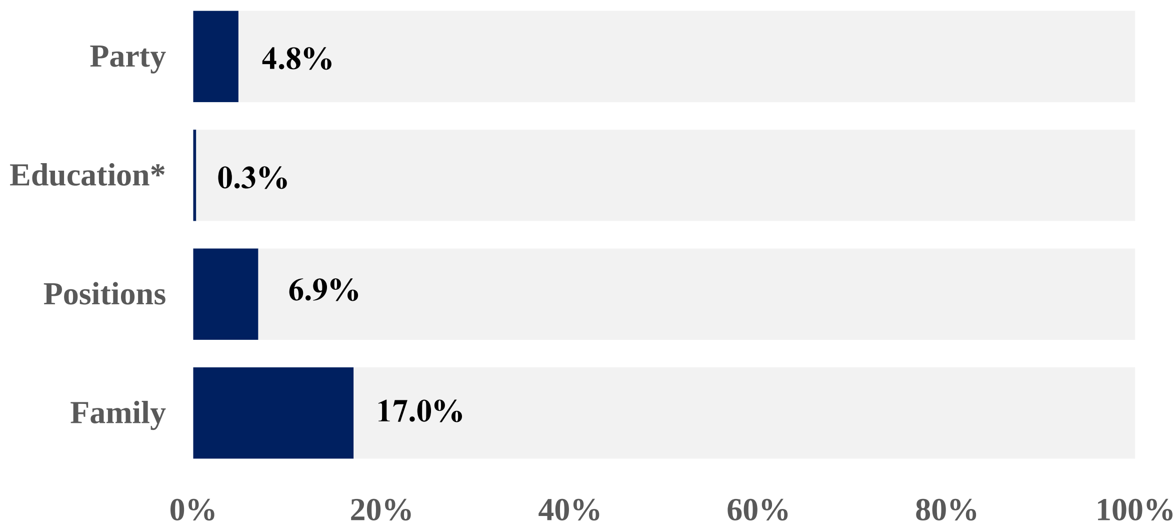
Figure 5. Proportion of rows that required manual cleaning in "multi-row attribute" tables



Some of these suspicious or defective rows could be easily fixed automatically through tweaks made to our original parsing code. While we can reduce the amount of manual cleaning however, some amount will always need to be done due to the non-uniformity of Wikipedia page templates. In any case, I attempted to calculate a hypothetical new diagnostic of how much of each table needed to be manually cleaned once I optimize and refine the original parsing code. As we can see below, with more refined and optimized parsing functions, we can cut down on the number of rows that require manual cleaning by a significant degree, particularly in the party and family tables:

[Figure 6 in next page]

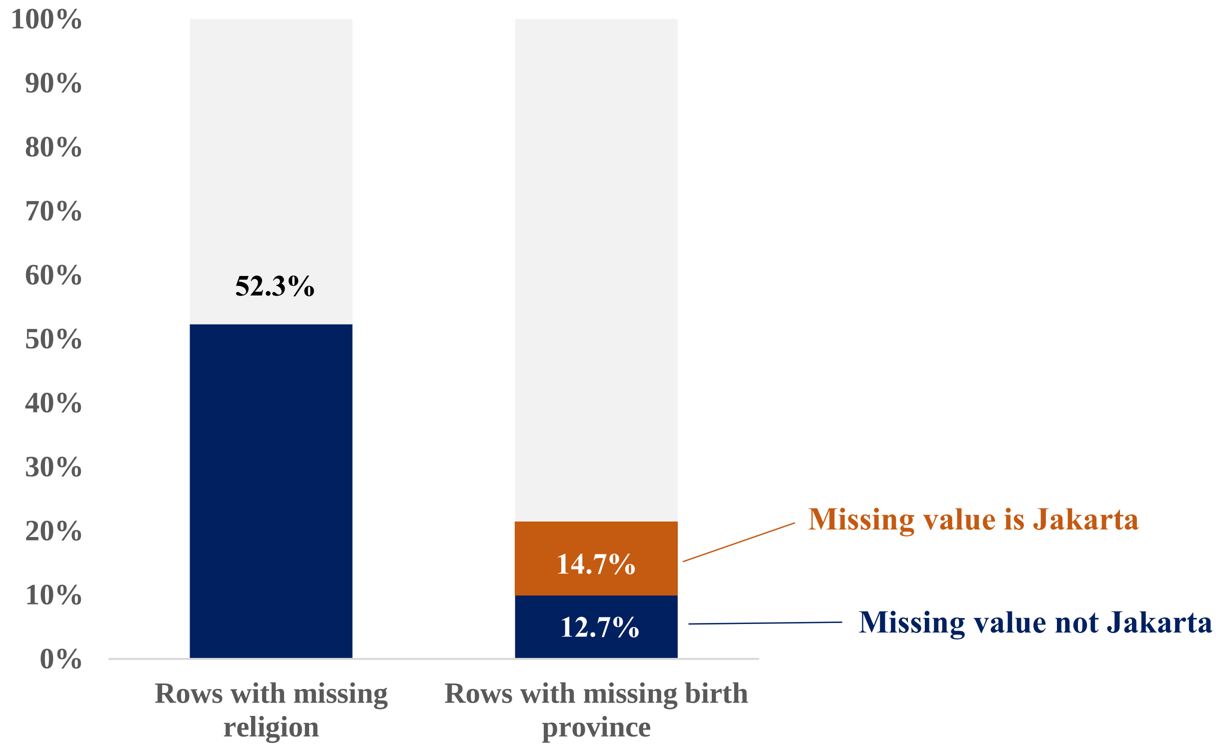
Figure 6. Expected proportion of rows that will require manual cleaning in "multi-row attribute" tables after refining parsing code



So far I've only provided a brief overview of the manual cleaning necessary and the extent to which improvements could potentially reduce them. For a more detailed rundown of all the manual checking done for each "multi-row attribute" table, as well as the assumed tweaks that would be necessary to produce the numbers seen in figure 6, please refer to **Appendix I**.

The manual checking done for the singulars table is quite different from that done for the above tables, mainly due to the fact that the fields extracted for this table tend to have a much more predictable structure in Wikipedia. As such, no big checking function was necessary. There were a few inconsistencies with how religion was parsed, but much of this was trivial and easily cleaned. The big challenge with regards to the singulars table pertained more to missingness, particularly in the province and religion fields.

Figure 7. Proportion of rows with missing values in the singulars table



Much of this missingness will have to be filled in with some manual imputation, of which I elaborate in further detail in the next segment on missingness. A small note to be made however, is that in a large portion of the rows with missing birth provinces, this missingness is due to my parsing code somehow not picking up Jakarta even when it appears in easily parse-able form in the extracted data. This is likely due to some still-unknown defect in the parsing code, and this portion of missing data should be filled once this issue is handled. Additionally, a lot of missing birth provinces are due to the parsing code not recognizing Dutch administrative names for contemporary Indonesian regions. [1] Once I accounted for this and Jakarta-related

[1] I parsed birth provinces by looking for matches between the extracted data on birth provinces and an alias table of all current provinces and regencies/cities in Indonesia. Since defunct Dutch administrative names are not included in this alias table, our parsing function was not able to detect them properly.

missingness, the proportion of rows with missing birth provinces goes down substantially to 6.8%.

On a final note, I should also mention that the largest chunk of my time doing “manual work” was spent in creating the alias tables that I used to translate each table. Fortunately, this is only “one-time” work in the sense that once we have a specific entity in an alias table already, we no longer need to create an entry for it again. To the extent that the entities that we have transcribed will repeat themselves when adding new entries to the database (this will certainly be the case for parties and positions), our existing alias tables should remain useful and applicable.

5.2. Missingness analysis

As mentioned in section 3.2 above, missingness was always going to be one of the biggest limitations of this dataset. The severity and type of missingness varies for each table, and will be outlined in further detail below:

- **Singulars:** As mentioned in section 5.1, there is significant missingness here with regards to religion and province (even after the code refinements mentioned in 5.1). There is little recourse here except to impute this by manually looking up other sources. Even after doing this however, it remains difficult to find this information for certain individuals, and a degree of missingness remains even after further research and manual imputation. Figure 8 below shows the degree of missingness before and after this manual search and imputation.
- **Party:** Since not everyone in the dataset is a politician, an accurate measure of missingness here would only look up how many politicians don’t have party information. [2] Even for those politicians who have no party membership, our dataset should ideally at least identify them

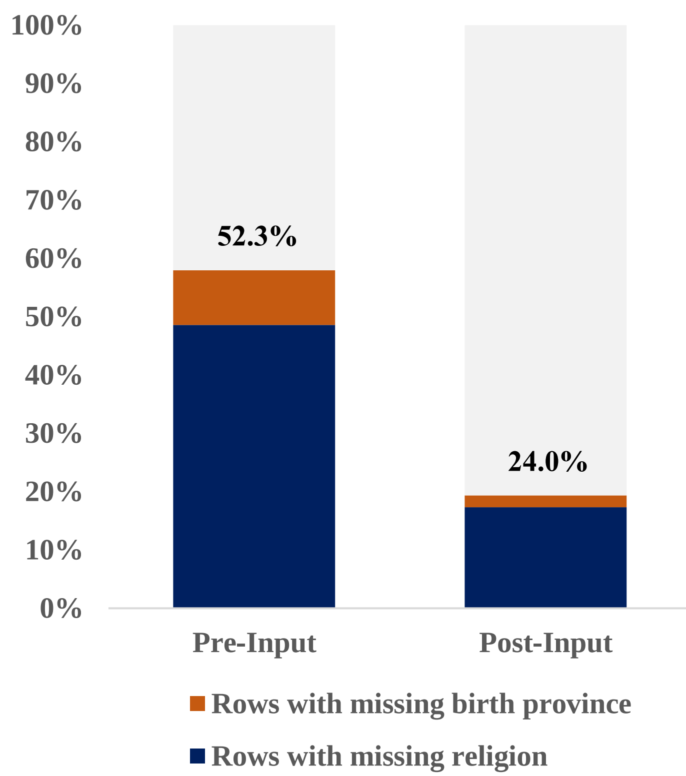
[2] Politicians here are defined as individuals with at least one position in the position table.

as independents. With these assumptions in mind, I found that a substantial number of politicians in our database – 57.2% to be precise – do not have their party information recorded.

- **Family:** Missingness here is trickier because it is difficult to distinguish between missing data and simply the absence of data. If we do not have information on person A's siblings, is it because this data is missing or simply because person A has no siblings? At the very least however, we know every person must have at least one family member (everyone must have a parent). With this minimum condition, we can approximate missingness by calculating how many persons in our database have at least one recorded relation. Using this calculation, I found that 45.1% of persons in our database have no recorded family relations at all.
- **Positions:** The missingness problem is also challenging here, mostly because conducting some form of missingness analysis itself is nigh impossible here. Indeed, unlike family, there is no minimum condition that stipulates one must hold at least a single political office in life. We can however, try to exclude the non-politicians for the moment and try to estimate some metric for the politicians in our database. Indeed, we mentioned previously in section 3.2 that 90.2% of Indonesia's cabinet members since 1945 have Wikipedia articles, and that this figure substantially improves to 99.4% when we only consider politicians post-1965. So we know for certain that this database at least captures the majority of cabinet-level politicians. What is still uncertain however, and remains difficult to ascertain, is whether our database records each individual politician's entire political career comprehensively.
- **Education:** It is perhaps here that the missingness problem is most intractable. Like political office, there is no minimum condition for education. Any proceeding missingness analysis here would have to assume that at least every politician in Indonesia attended some institute

of higher education (since most of the educational information recorded by Wikipedia focuses on higher education), which is a tenuous assumption at best. As such, we remain largely in the dark as to the exact degree of missingness in this table.

Figure 8. Proportion of rows in singulars table with missing values pre and post-manual imputation to the missing values



Unfortunately, it appears that the problem of missingness might be difficult to resolve for databases generated by our Wikipedia-based method. While manual imputation is certainly an option for relatively straightforward fields such as religion and place of birth, this is trickier for more complex fields such as positions. Even if this information was available, the amount of manual work needed to fill in these values invalidates the very cost and time-effectiveness that is the primary motivating factor for this method. Furthermore, it appears to be the case that much of the missingness observed is also not random. More famous and influential figures such as Presidents tend to have more complete information on them compared to more obscure politicians. As such, simply dropping missing values risks biasing the data in a way that disproportionately weights the data towards politicians with greater star power. Indeed, this risks ignoring those very *eminences grises* that served as a key motivating factor for this study in the first place. Additionally, a lot of this missingness is "unknown" in the sense that we often don't know whether a missing value reflects genuine missingness or simply the absence of a phenomenon. The specter of missingness then, needs to loom large over the heads of every responsible researcher who decides to make use of this database.

6. Conclusion

Overall, it appears that our method has been able to generate a dataset that captures salient information on Indonesia's political elite. The coverage ratio is quite high, with the database containing records on 90.2% of all ministers who have served in Indonesia since its founding in 1945, as well as a substantial number of additional records on familial relations to these very same ministers (these additional familial relations make up 34.7% of all the persons recorded in the database). Sections 4.2 and 4.3 above also demonstrate the various kinds of analysis that one can reasonably do using this database. All this demonstrates that it is indeed possible to generate a useful and working database on political elites with a scraping-based computational method from public sources such as Wikipedia. The general method used here should also be replicable for any attempt to generate a similar database for other countries, and much of the manual cleaning involved is relatively straightforward and would require at most some basic guidelines. Even if the fine details of how information should be parsed may differ, the general method outlined here should also be applicable for generating a similar database from other online sources beyond Wikipedia. Finally, it should also be more than doable to expand this database to include non-cabinet political elites such as military personnel or legislative members. Indeed, one of the original goals of this study was to create a database that had a broader definition of political elites beyond just cabinet members. Having demonstrated the veracity of the generative method outlined here, applying this method to expand this database beyond cabinet members should be relatively straightforward.

However, there are some limitations that need to be kept in mind. While the initial goal was to develop a fully automated workflow for generating the database, the irregular and unstructured nature of Wikipedia-based data means that some degree of manual cleaning is likely still necessary to account for unexpected structures and formats. It is likely then, that the workflow for similarly generated databases in the future will end up featuring some hybrid of automated extraction and parsing, as well as some manual cleaning and validation.

A more pressing structural flaw however, is data missingness. Indeed, short of fully adopting expert-centric and journalistic or interview-based techniques of data generation, it is likely that some degree of missingness will always be present in databases generated from "crowdsourced" stores of information like Wikipedia. Porter, Verdery and Gaddis' (2020) three V's come to mind here, particularly big data's tendency to fall short in the realms of validity and value. The extent to which this missingness will be a problem is largely dependent on the quality of the original data source, something which might vary with different linguistic versions of Wikipedia. It is entirely plausible that information on say US politicians, extracted from the English version of Wikipedia, may be much more well-structured and comprehensive than its

Indonesian counterpart. It is important to reiterate then, that full disclosure of missingness is critical in any dissemination of databases generated by this method.

These various considerations point to several avenues of further research within this area. First, further information can still be mined out of Wikipedia. Indeed, this study focuses solely on extracting data from infoboxes, but a great deal of information is also stored in the main text of Wikipedia pages. Some of this information even has a degree of structure that makes it amenable to the regular expression-based parsing method used here. For instance, the educational history of various politicians is frequently found as part of a bulleted list in the main body of their Wikipedia pages. These lists should have a degree of regularity that can be exploited by regular expression-based functions. This also brings back to mind the problem of missingness, since a lot of the missing educational information in the generated database is clearly due to our parsing function not picking up this data if it is located in the main text rather than the infobox.

Second, exploring alternative parsing methods outside regular expressions can also be promising. Researchers such as Beavan and Nanni (2021) and Dai, Olah and Le (2015) have shown the merits of using NLP-based methods to extract information from various social science texts. The initial plan was in fact to implement similar methods to parse information from Wikipedia. Unfortunately, the lack of NLP models in the Indonesian language made this option infeasible for this study. The potential remains however, and development of multilingual NLP models would be a welcome development, especially since any attempt to expand empirical information on elites outside the developed world will likely rely largely on information from non-English sources. On a similar note, the advent of large language models (LLMs) also provides an interesting avenue of exploration. Suitable NLP or LLM-based methods have the advantage of being able to parse and extract information from completely unstructured texts, allowing us to expand beyond simple infoboxes to include truly unstructured texts in a way that regular expression-based methods simply cannot handle.

Finally, no discussion of this generative method will be complete without a consideration of original data sources. Ultimately, any database generated by scraping-based methods will only be as good as its original data source. For all of its strengths, one of the most glaring holes in our database is its lack of information on informal relationships and power structures. Indeed, few politico-business relations or explicit non-family relations seem to have made their way to the final database. This paper started by acknowledging the importance of piercing the veil of formal politics, and this exercise has shown that Wikipedia can only pierce this veil so far. It is in this regard that the classical, expert-based methods of data generation will still prove to be invaluable. And it bears repeating as well that the missingness inherent to such "crowdsourced" datasets means that any formal analysis based on them will have to be conducted with a careful and skeptical eye. Where such datasets prove to be very useful however – and this is where missingness is less pressing an issue and their breadth really shines – is in exploratory work. We saw this to some degree in sections 4.2 and 4.3 above, where even some relatively simple exercises were able to generate interesting insights that, while certainly not definitive, could serve as launchpads for deeper, more comprehensive studies. The most useful approach going forward then, is to see these different methods as broadly complementary. The stereotype of quality versus quantity is a well-worn one, but it holds true in this case. Combined, we have the ability to take advantage of both: the speed and breadth of big data, paired with the depth of domain knowledge; the wisdom of the crowd, paired with the wisdom of the expert.

References

Finley, M. I. “Athenian Demagogues.” *Past and Present*, 21(1): 3-24, April 1962. https://doi.org/10.1093/past/21.1.3

Mosca, Gaetano. *The Ruling Class*. Translated by Hannah D. Kahn. New York and London: McGraw-Hill Book Company Inc., 1939.

Bottomore, Tom. *Elites and Society*. London: C. A. Watts & Co. Ltd., 1964.

Dahl, Robert A. “A Critique of the Ruling Elite Model.” *The American Political Science Review*, 52(2): 463-469, June 1958. https://doi.org/10.2307/1952327

Mills, C. Wright. *The Power Elite*. New York: Oxford University Press, 1959.

Zuckerman, Alan. “The Concept ‘Political Elite’: Lessons from Mosca and Pareto.” *The Journal of Politics*, 39(2): 324-344, May 1977. https://doi.org/10.2307/2130054

Beteille, Andre. “Institutions and networks.” *Current Science*, 97(2): 148-156, July 2009. https://www.jstor.org/stable/24111911

Gerring, J., Oncel, E., Morrison, K., Keefer, P. “The Global Leadership Project: A Comprehensive Database of Political Elites.” September 2014. https://dx.doi.org/10.2139/ssrn.2491672

Shih, V., Meyer, D., Lee, J. “Factions of Different Stripes: Gauging the Recruitment Logics of Factions in the Reform Period.” *Journal of East Asian Studies*, 16(1): 43-60, March 2016. https://www.jstor.org/stable/26335169

Jiang, Junyan. “Making Bureaucracy Work: Patronage Networks, Performance Incentives, and Economic Development in China.” *American Journal of Political Science*, 62(4): 982-999, October 2018. https://www.jstor.org/stable/26598796

Best, H., Edinger, M. “Converging Representative Elites in Europe? An Introduction to the EurElite Project.” *Sociologicky Casopsis / Czech Sociological Review*, 41(3): 499-510, June 2005. https://www.jstor.org/stable/41132162

Dowding, K., Dumont, P. *The Selection of Ministers in Europe; Hiring and Firing*. Routledge, 2009.

Aspinall, E., Fossati, D., Muhtadi, B., Warburton, E. “Mapping the Indonesian Political Spectrum.” *New Mandala*. Web. Last updated 24 Apr, 2018. https://www.newmandala.org/mapping-indonesian-political-spectrum/

Rusnaedy, Z. & Purwaningsih, T. “Keluarga politik Yasin Limpo pada pemilihan kepala daerah di Kabupaten Gowa tahun 2014,” *Jurnal Politik*, 3(2): 5, https://doi.org/10.7454/jp.v3i2.116

Burchanuddin, A., Adam, A., Alim, A., Agustang, A. “Cultural Reproduction in the Socio-political Context of Bone District, South Sulawesi, Indonesia.” *The Journal of Sociology & Social Welfare*, 12(1-2): 12-22, March 2021. http://dx.doi.org/10.31901/24566764.2021/12.1-2.361

Muksin, D., Purwaningsih, T., Nurmandi, A. “Praktik dinasti politik di aras lokal pasca Reformasi: Studi kasus Abdul Gani Kasuba dan Ahmad Hidayat Mus pada Pilkada provinsi Maluku Utara.” *Jurnal Wacana Politik*, 4(2): 133-144, 2019. https://doi.org/10.24198/jwp.v4i2.25336

Kelihu, Ardiman. “Political Clientelism, Family Power and Conflict Permanence in Pilkada; The Case From Maluku.” *Power Conflict Democracy Journal*, 10(1): 75-108, November 2022. https://doi.org/10.22146/pcd.v10i1.5417

Chalik, A., & Latif, M. “Exploring the formation of coalitions between Islamist and secular parties in Indonesia local elections: Figure, patronage and common enemy.” *Hamdard Islamicus*, 43(2): 1761-1765, 2020.

Nyrup, J., & Bramwell, S. “Who Governs? A New Global Dataset on Members of Cabinets.” *Cambridge University Press*, 114(4): 1366-1374, November 2020. https://doi.org/10.1017/S0003055420000490

Porter, N., Verdery, A., Gaddis, S. “Enhancing big data in the social sciences with crowdsourcing: Data augmentation practices, techniques and opportunities.” *PLoS ONE*, 15(6): e0233154, June 2020. https://doi.org/10.1371/journal.pone.0233154

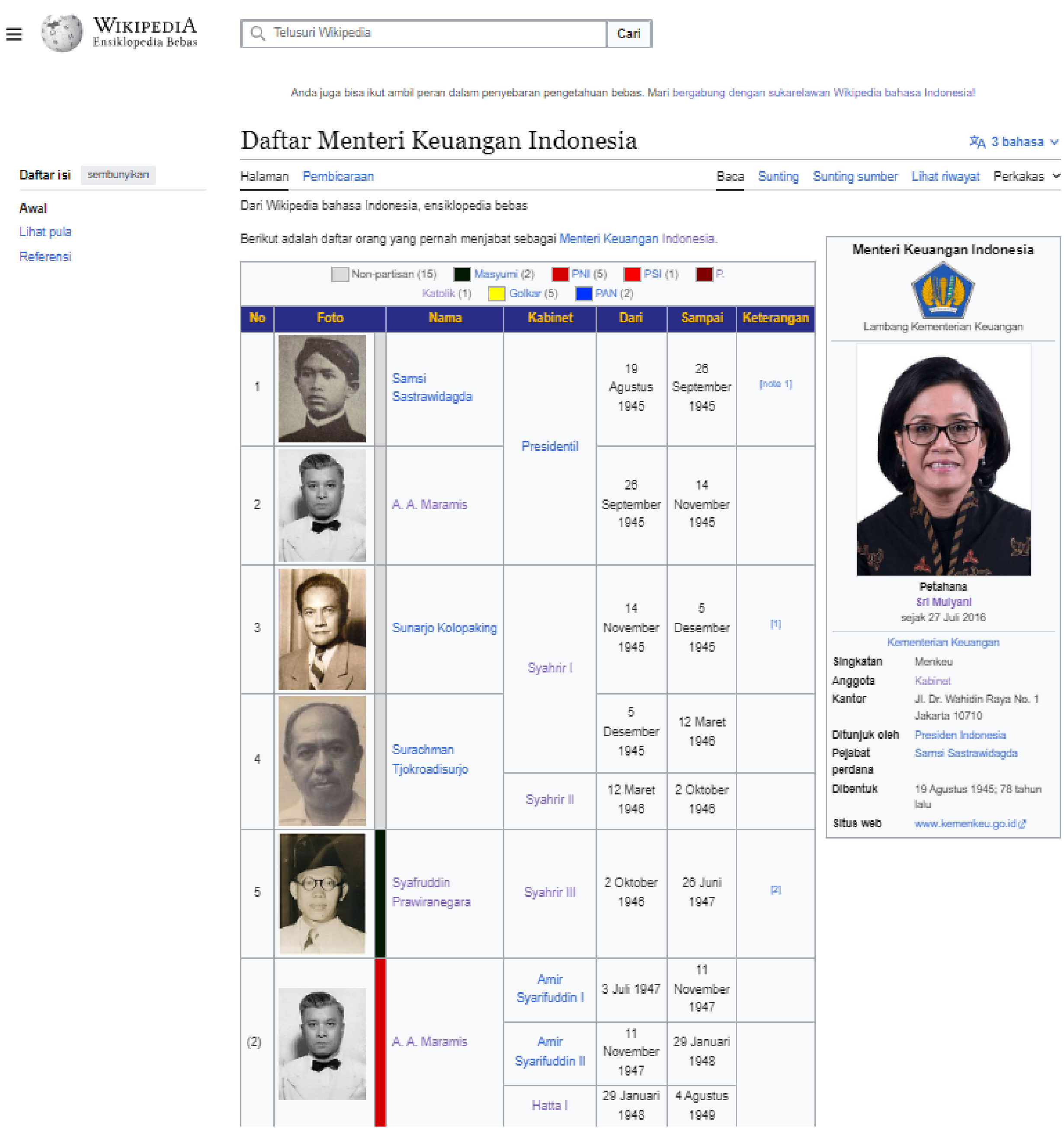
Razak, Imanuddin. “Minang supremacy: A declining triumphalism.” *The Jakarta Post*. Web. Last updated Feb 10, 2020. https://www.thejakartapost.com/news/2020/02/10/minang-supremacy-a-declining-triumphalism.html

Beavan, D., & Nanni, F. “Data study group final report: The National archives, UK; Discovering topics and trends in the UK government web archive.” *Zenodo*, 2021. https://doi.org/10.5281/zenodo.4981184

Dai, A., Olah, C., Le., Q. “Document Embedding with Paragraph Vectors.” *arXiv*: 1507.07998, July 2015. https://doi.org/10.48550/arXiv.1507.07998

APPENDICES

Appendix A: Sample list page from Wikipedia



WIKIPEDIA
Ensiklopedia Bebas

Telusuri Wikipedia | Cari

Anda juga bisa ikut ambil peran dalam penyebaran pengetahuan bebas. Mari bergabung dengan sukarelawan Wikipedia bahasa Indonesia!

Daftar isi sembunyikan

Awal
Lihat pula
Referensi

Daftar Menteri Keuangan Indonesia

3 bahasa

Halaman Pembicaraan | Baca Sunting Sunting sumber Lihat riwayat Perkakas

Dari Wikipedia bahasa Indonesia, ensiklopedia bebas

Berikut adalah daftar orang yang pernah menjabat sebagai Menteri Keuangan Indonesia.

Non-partisan (15) Masyumi (2) PNI (5) PSI (1) P. Katolik (1) Golkar (5) PAN (2)

No	Foto	Nama	Kabinet	Dari	Sampai	Keterangan
1		Samsi Sastrawidagda	Presidentil	19 Agustus 1945	26 September 1945	[note 1]
2		A. A. Maramis		26 September 1945	14 November 1945	
3		Sunarjo Kolopaking	Syahrir I	14 November 1945	5 Desember 1945	[1]
4		Surachman Tjokroadisurjo		5 Desember 1945	12 Maret 1946	
			Syahrir II	12 Maret 1946	2 Oktober 1946	
5		Syafruddin Prawiranegara	Syahrir III	2 Oktober 1946	26 Juni 1947	[2]
(2)		A. A. Maramis	Amir Syarifuddin I	3 Juli 1947	11 November 1947	
			Amir Syarifuddin II	11 November 1947	29 Januari 1948	
			Hatta I	29 Januari 1948	4 Agustus 1949	

Menteri Keuangan Indonesia

Lambang Kementerian Keuangan

Petahana
Sri Mulyani
sejak 27 Juli 2016

Kementerian Keuangan

Singkatan	Menkeu
Anggota	Kabinet
Kantor	Jl. Dr. Wahidin Raya No. 1 Jakarta 10710
Ditunjuk oleh	Presiden Indonesia
Pejabat perdana	Samsi Sastrawidagda
Dibentuk	19 Agustus 1945; 78 tahun lalu
Situs web	www.kemenkeu.go.id

Appendix B: Sample raw Wikipedia infobox data

Here are a few samples of the raw, uncleaned data extracted from the Wikipedia API.

This is what was extracted from the “children” field in Megawati Soekarnoputri’s page:

```
[[Mohammad Rizki Pratama]] <br/> [[Mohammad Prananda Prabowo]] <br/><small> (dari [[Surindro Supjarso]]) </small> <br /> [[Puan Maharani]] <br/><small>(dari [[Taufiq Kiemas]])</small>
```

This is what was extracted from the “spouse” field in Sudomo’s page:

```
{{marriage|Fransisca Play|1961|1980|reason=divorced}} {{marriage|[[Siska Widowati|Fransisca Diah Widhowaty]]|20 September 1990|1994|reason=divorced}} {{marriage|Aty Kusumawaty|1998|2002|reason=divorced}}
```

This is what was extracted from the “children” field in Edi Sudradjat’s page:

```
{{unbulleted list|1. [[Insinyur|Ir.]] Iwan Darmawan|2. [[Insinyur|Ir.]] Ita Setia Wati|3. [[Iman Budiman|Brigjen TNI (Purn.) Iman Budiman]]|4. [[Andi Gunawan|Kolonel Inf. Andi Gunawan]]}}
```

Appendix C: Database Metadata

Singulars table

Field	Description
name	A specific person's name. Should be unique in this table.
Pageid	The ID of this person's Wikipedia page. Should be unique in this table.
Wikidata	This person's unique Wikidata ID. Will be used as a key to connect with other instances of this person in the other tables.
Dob	This person's year of birth. Missing values are represented by a 0.
Death	This person's year of death. Missing values are represented by a 0.
Religion	This person's recorded religion. At the moment, this cannot handle multiple religions (in the case that an individual switches religions), so this field defaults to recording a person's latest religion.
Province	This person's province of birth.
Region	This person's regency/city of birth.
Area	This person's area of birth (essentially the macro regions Java, Sumatra, Greater Jakarta, Kalimantan, Sulawesi, Bali-Nusra and "Maluku and Papua.")
born_abroad	A Boolean value that records whether this individual was born abroad or not

Positions table

Field	Description
person	A specific person's name.
personid	The unique Wikidata ID of this person. This is the same as the 44ikidata field in the singulars table.
Position	A specific office/position this person is (or was once) holding
position_id	The unique Wikidata ID of this position.
Start	The year this person started holding this position. Missing values are represented by a 0.
End	The year this person stopped holding this position. Missing values are represented by a 0.
Type	The type of this position (eg. Executive, Military, Legislative, NGO, etc.)
country	The country this position is associated with. So if the position in this row is "Minister of Defense" and the country is "Indonesia," the position refers to the "Minister of Defense of Indonesia."

Party table

Field	Description
person	A specific person’s name.
personid	The unique Wikidata ID of this person. This is the same as the 45ikidata field in the singulars table.
Party	A party the person is (or was) a member of.
Party_id	The unique Wikidata ID of this party.

Education table

Field	Description
person	A specific person’s name.
personid	The unique Wikidata ID of this person. This is the same as the 45ikidata field in the singulars table.
Educ	An educational institution this person is (or once) attending.
Edu_id	The unique Wikidata ID of this educational institution.
Country	The country the educational institution in this row is located in. The values “Indonesia (Dutch)” and “Indonesia (Japan)” refer to Dutch or Japanese institutions during their periods of rule in Indonesia.
Pesantren_madrasah	A Boolean value that specifies whether this institution is a pesantren or madrasah or not.
Military_police	A Boolean value that specifies whether this institution is a military or police academy or not.

[Appendix continues in the next page]

Family table

Field	Description
person	A specific person's name.
personid	The unique Wikidata ID of this person. This is the same as the wikidata field in the singulars table.
relative	Another person who is somehow related to the person in the "person" field.
rel_id	The unique Wikidata ID of the person in the "relative" field.
status	The relationship status between "person" and "relative." For example, if status here is "uncle," then "person" is "relative's" uncle.
distance_birth	Counts the relational distance between "person" and "relative" by birth. Only birth and spousal relations count as 1. For example, a father or spouse has a distance of 1. A sibling however, has a distance of 2 since the birth connection needs to go through a parent first.
distance_nuclear	Counts the relational distance between "person" and "relative" by nuclear family. Distance is counted as 1 here as long as two individuals are part of the same nuclear family. Unlike the above then, a sibling has a distance of 1 since both are part of the same nuclear family.
family	A Boolean function that specifies whether this relationship is familial or not. Included to account for the few non-familial relationships recorded here (eg. student, colleague).

Appendix D: Screenshot of singulars table

name	pageid	wikidata	dob	death	religion	province	region	area	born_abroad
Abikusno Tjokrosujoso	1174515	Q4667741	1897	1968	Islam	Central Java	Tegal	Java	F
Adnan Kapau Gani	205654	Q4264085	1905	1968	Islam	West Sumatra	Agam	Sumatra	F
Ali Sastroamidjojo	48386	Q2669326	1903	1975	Islam	Central Java	Magelang	Java	F
Djody Gondokusumo	3226465	Q109430802	1912	0	Islam	Yogyakarta	Yogyakarta	Java	F
Fatmawati	7082	Q468519	1923	1980	Islam	Bengkulu	Bengkulu	Sumatra	F
Ferdinand Lumban Tobing	68191	Q11188587	1899	1962	Protestant	North Sumatra	Central Tapanuli	Sumatra	F
Guruh Soekarnoputra	196541	Q4264069	1953	0	Islam	Jakarta		Greater Jakarta	F
Hazairin	7093	Q11168814	1906	1975	Islam	West Sumatra	Bukittinggi	Sumatra	F
Iskak Tjokroadisurjo	2494668	Q56392413	1896	1984		East Java	Jombang	Java	F
Iwa Kusumasumantri	48946	Q10975733	1899	1971	Islam	West Java	Ciamis	Java	F
Joko Widodo	31706	Q3318231	1961	0	Islam	Central Java	Surakarta	Java	F
Lie Kiat Teng	33036	Q12494804	1912	1983	Islam	West Java	Sukabumi	Java	F
Ma'ruf Amin	344316	Q12497177	1943	0	Islam	Banten	Tangerang	Java	F
Masjkur	345021	Q12497345	1904	1994	Islam	East Java	Malang	Java	F
Megawati Soekarnoputri	3483	Q76179	1947	0	Islam	Yogyakarta	Yogyakarta	Java	F
Mohammad Hanafiah	1414101	Q17410754	1904	1981	Islam	South Kalimantan		Kalimantan	F
Mohammad Hatta	7125	Q29050	1902	1980	Islam	West Sumatra	Bukittinggi	Sumatra	F
Mohammad Yamin	7126	Q3503054	1903	1962	Islam	West Sumatra	Sawahlunto	Sumatra	F
Ong Eng Die	234809	Q7093799	1910	0	Catholic	Gorontalo	Gorontalo	Sulawesi	F
Pandji Suroso	256708	Q129689	1893	1981	Islam	East Java	Sidoarjo	Java	F
R. Sunarjo	3001391	Q85993732	1908	1996	Islam	Central Java	Sragen	Java	F
Rachmawati Soekarnoputri	371014	Q12507600	1950	2021	Islam	Jakarta		Greater Jakarta	F
Ratna Sari Dewi Soekarno	255063	Q4263050	1940	0		Japan			T
Roosseno Soerjohadikoesoemo	302490	Q12508921	1908	1996	Islam	East Java	Madiun	Java	F
Sadjarwo Djarwonagoro	3080950	Q95544874	1917	1996		Central Java	Surakarta	Java	F
Soedibjo	1488546	Q17411444	1918	2008	Islam	East Java	Probolinggo	Java	F
Soekarno	2834	Q76127	1901	1970	Islam	East Java	Surabaya	Java	F
Sukmawati Soekarnoputri	325775	Q12517042	1951	0	Hindu	Jakarta		Greater Jakarta	F
Sunario Sastrowardoyo	157808	Q4446092	1902	1997	Islam	East Java	Madiun	Java	F
Sutan Muchtar Abidin	1026825	Q12518231	1910	0	Islam	West Sumatra	Pariaman	Sumatra	F
Wongsonegoro	346330	Q3526552	1897	1978	Kejawen	Central Java	Surakarta	Java	F
Zainul Arifin Pohan	11479	Q12525401	1909	1963	Islam	North Sumatra	Central Tapanuli	Sumatra	F
Agustinus Suhardi	2741652	Q61714760	1899	0	Catholic	Central Java	Klaten	Java	F
Dahlan Ibrahim	1017090	Q12480685	0	0	Islam				F
Djuanda Kartawidjaja	7101	Q2670453	1911	1963	Islam	West Java	Tasikmalaya	Java	F
Eny Karim	1017375	Q16044649	1910	1995	Islam	East Java	Batu	Java	F
Handrianus Sinaga	2122676	Q28723855	1912	1981	Protestant	North Sumatra	Samosir	Sumatra	F
Idham Chalid	79382	Q6854752	1921	2010	Islam	South Kalimantan	Tanah Bumbu	Kalimantan	F
Jusuf Wibisono	345024	Q6854709	1909	1982	Islam	Central Java	Magelang	Java	F
Moeljatno	831715	Q6890147	1909	1971	Islam	Central Java	Surakarta	Java	F
Mohammad Roem	27480	Q2746664	1908	1983	Islam	Central Java	Temanggung	Java	F
Pangeran Mohammad Nur	116336	Q12502601	1901	1979	Islam	South Kalimantan		Kalimantan	F
Roeslan Abdulgani	24875	Q967941	1914	2005	Islam	East Java	Surabaya	Java	F
Rusli Abdul Wahid	1027645	Q12509202	1908	1999	Islam	West Sumatra	Lima Puluh Kota	Sumatra	F
Sabilal Rasjad	1026653	Q12511399	1908	0	Islam	West Sumatra	Agam	Sumatra	F
Sarino Mangunpranoto	346335	Q12512234	1910	1983	Islam	Central Java	Purworejo	Java	F
Wirjono Prodjodikoro	437179	Q8026958	1903	1985	Islam	Central Java	Surakarta	Java	F
AA Maramis	49504	Q12682432	1897	1977	Protestant	North Sulawesi	Manado	Sulawesi	F
Achmad Asj'ari	824093	Q14404403	0	0	Islam	South Sumatra	Palembang	Sumatra	F
Agus Salim	7062	Q118629	1884	1954	Islam	West Sumatra	Agam	Sumatra	F
Amir Sjarifuddin	41020	Q1810083	1907	1948	Protestant	North Sumatra	Medan	Sumatra	F
Anwaruddin	824101	Q4778076	0	0	Islam				F
Arudji Kartawinata	202142	Q13198631	1905	1970	Islam	West Java	Garut	Java	F
Hamengkubuwono IX	792	Q76227	1912	1988	Islam	Yogyakarta	Yogyakarta	Java	F
Herling Laoh	661543	Q6696761	1902	1970	Protestant	North Sulawesi	Minahasa	Sulawesi	F
IJ Kasimo	211023	Q13534328	1900	1986	Catholic	Yogyakarta	Yogyakarta	Java	F
J. Leimena	48756	Q2642883	1905	1977	Protestant	Maluku	Ambon	Maluku and Papua	F
SK Trimurti	140757	Q7387595	1912	2008		Central Java	Boyolali	Java	F
Satrio	433006	Q12512336	1916	1986	Islam	East Java	Banyuwangi	Java	F

Appendix E: Screenshot of positions table

person	personid	position_id	start	end	position	type	country
A.M. Hendropriyono	Q4666027	Q25466621	1998	1998	Minister of Transmigration	Executive	Indonesia
A.M. Hendropriyono	Q4666027	Q25453522	2001	2004	Director of the State Intelligence Agency	Intelligence	Indonesia
A.M. Hendropriyono	Q4666027	Q13405290	2016	2018	Chairman of the Indonesian Justice and Unity Party (PKPI)	Party	Indonesia
A.R. Soehoed	Q12470435	Q19725118	1978	1983	Minister of Industry	Executive	Indonesia
AA Maramis	Q12682432	Q4434739	1948	1949	Minister of Foreign Affairs	Executive	Indonesia
AA Maramis	Q12682432	Q3212427	1945	1945	Minister of Finance	Executive	Indonesia
AA Maramis	Q12682432	Q65212583	1950	1953	Ambassador to the Philippines	Diplomatic	Indonesia
AA Maramis	Q12682432	Q31174404	1953	1956	Ambassador to West Germany	Diplomatic	Indonesia
AA Maramis	Q12682432	Q85989675	1956	1959	Ambassador to the Russian Federation	Diplomatic	Indonesia
AA Maramis	Q12682432	Q85989707	1958	1960	Ambassador to Finland	Diplomatic	Indonesia
Abdoel Halim	Q2978474	Q672635	1950	1950	Prime Minister	Executive	Indonesia
Abdoel Halim	Q2978474	Q11046530	1950	1951	Minister of Defense	Executive	Indonesia
Abdul Gafur (politikus)	Q12470547	Q19725154	1997	1999	Deputy Speaker of the People's Consultative Assembly	Legislative	Indonesia
Abdul Gafur (politikus)	Q12470547	Q12479876	1978	1988	Minister of Youth and Sports Affairs	Executive	Indonesia
Abdul Hakim Harahap	Q19942211	Q12479942	1950	1950	Deputy Prime Minister	Executive	Indonesia
Abdul Hakim Harahap	Q19942211	Q12479848	1951	1953	Governor of North Sumatra	Regional	Indonesia
Abdul Hakim Harahap	Q19942211	Q2670027	1956	1960	Member of the House of Representatives	Legislative	Indonesia
Abdul Hakim Harahap	Q19942211	O14	1955	1956	Minister of State (Defense)	Executive	Indonesia
Abdul Halim Iskandar	Q72061441	Q12479874	2019	0	Minister of Villages, Disadvantaged Regions, and Transmigration	Executive	Indonesia
Abdul Haris Nasution	Q317291	Q11046530	1959	1966	Minister of Defense	Executive	Indonesia
Abdul Haris Nasution	Q317291	Q12479854	1966	1972	Speaker of the People's Consultative Assembly	Legislative	Indonesia
Abdul Haris Nasution	Q317291	Q11281667	1955	1959	Commander of the Indonesian National Armed Forces	Military	Indonesia
Abdul Haris Nasution	Q317291	Q14917366	1949	1952	Chief of Staff of the Indonesian Army	Military	Indonesia
Abdul Haris Nasution	Q317291	Q75137905	1948	1953	Deputy Commander of the Indonesian National Armed Forces	Military	Indonesia
Abdul Latief (pengusaha)	Q4665497	Q19725113	1993	1998	Minister of Manpower	Executive	Indonesia
Abdul Latief (pengusaha)	Q4665497	Q12479872	1998	1998	Minister of Culture and Tourism	Executive	Indonesia
Abdul Latif Amin Imron	Q107994307	Q20426412	2018	2022	Regent of Bangkalan	Regional	Indonesia
Abdul Malik Fadjar	Q11109918	Q12479877	2001	2004	Minister of Education	Executive	Indonesia
Abdul Malik Fadjar	Q11109918	Q4272757	2015	2019	Member of the Presidential Advisory Council	Advisory	Indonesia
Abdul Malik Fadjar	Q11109918	Q12479860	1998	1999	Minister of Religious Affairs	Executive	Indonesia
Abdul Malik Fadjar	Q11109918	Q13095079	2004	2004	Coordinating Minister of Human and Cultural Development	Executive	Indonesia
Abdul Rahman Saleh (jaksa)	Q4665652	Q16179573	2000	2004	Justice of the Supreme Court of Indonesia	Judicial	Indonesia
Abdul Rahman Saleh (jaksa)	Q4665652	Q56388152	2004	2007	Attorney General of Indonesia	Executive	Indonesia
Abdul Rahman Saleh (jaksa)	Q4665652	Q85989631	2008	2011	Ambassador to Denmark	Diplomatic	Indonesia
Abdul Syukur	Q109433339	Q25461131	2009	2014	Member of the Banten Regional House of Representatives	Regional legislati	Indonesia
Abdul Syukur	Q109433339	Q25461133	1999	2009	Member of the Tangerang City Regional House of Representatives	Regional legislati	Indonesia
Abdullah Aidit	Q108782375	Q2670027	1950	1954	Member of the House of Representatives	Legislative	Indonesia
Abdullah Amu	Q12470573	Q14919621	1966	1967	Governor of North Sulawesi	Regional	Indonesia
Abdullah Azwar Anas	Q16162595	Q19725115	2022	0	Minister of State Apparatus Utilization and Bureaucratic Reform	Executive	Indonesia
Abdullah Azwar Anas	Q16162595	Q4121272	2022	2022	Chief of the Institute of Procurement Policy	Agency	Indonesia
Abdullah Azwar Anas	Q16162595	Q16173194	2016	2021	Regent of Banyuwangi	Regional	Indonesia

Appendix F: Screenshot of party table

person	personid	party_id	party
Abikusno Tjokrosujoso	Q4667741	Q4200848	Indonesian Sarikat Islam Party
Djody Gondokusumo	Q109430802	Q4314852	National People's Party
Guruh Soekarnoputra	Q4264069	Q2084109	Indonesia Democratic Party (PDI)
Guruh Soekarnoputra	Q4264069	Q1186306	Indonesian Democratic Party of Struggle (PDI-P)
Iskak Tjokroadisurjo	Q56392413	Q1965221	Indonesian National Party (PNI)
Joko Widodo	Q3318231	Q1186306	Indonesian Democratic Party of Struggle (PDI-P)
Lie Kiat Teng	Q12494804	Q4200848	Indonesian Sarikat Islam Party
Ma'ruf Amin	Q12497177	F1	Independent
Masjkur	Q12497345	Q686441	Nahdlatul Ulama
Megawati Soekarnoputri	Q76179	Q2084109	Indonesia Democratic Party (PDI)
Megawati Soekarnoputri	Q76179	Q1186306	Indonesian Democratic Party of Struggle (PDI-P)
Mohammad Hatta	Q29050	F1	Independent
Mohammad Yamin	Q3503054	Q19730387	Indonesian Party
Mohammad Yamin	Q3503054	Q110413286	Indonesian People's Movement
R. Sunarjo	Q85993732	Q2669302	United Development Party (PPP)
Rachmawati Soekarnoputri	Q12507600	Q7196857	Pioneers' Party
Rachmawati Soekarnoputri	Q12507600	Q4207219	NasDem Party
Rachmawati Soekarnoputri	Q12507600	Q4261459	Gerindra Party
Soedibjo	Q17411444	Q4200848	Indonesian Sarikat Islam Party
Soekarno	Q76127	Q1965221	Indonesian National Party (PNI)
Sukmawati Soekarnoputri	Q12517042	Q4261536	Indonesian National Marhaenist Party
Sunario Sastrowardoyo	Q4446092	Q1965221	Indonesian National Party (PNI)
Sutan Muchtar Abidin	Q12518231	Q116299110	Labor Party (1998)
Dahlan Ibrahim	Q12480685	Q12503305	Indonesian Independence Supporters Movement Party
Eny Karim	Q16044649	Q1965221	Indonesian National Party (PNI)
Handrianus Sinaga	Q28723855	Q4200847	Indonesian Christian Party (Parkindo)
Idham Chalid	Q6854752	Q2579297	Masyumi Party
Idham Chalid	Q6854752	Q686441	Nahdlatul Ulama
Idham Chalid	Q6854752	Q2669302	United Development Party (PPP)
Jusuf Wibisono	Q6854709	Q2579297	Masyumi Party
Moeljatno	Q6890147	Q2579297	Masyumi Party
Rusli Abdul Wahid	Q12509202	Q12505151	Tarbiyah Islamiyah Union
Rusli Abdul Wahid	Q12509202	Q2669302	United Development Party (PPP)
AA Maramis	Q12682432	Q1965221	Indonesian National Party (PNI)
Amir Sjarifuddin	Q1810083	Q204699	Indonesian Socialist Party (PSI)
Anwaruddin	Q4778076	Q4200848	Indonesian Sarikat Islam Party
IJ Kasimo	Q13534328	Q5053222	Indonesia Catholic Party
J. Leimena	Q2642883	Q4200847	Indonesian Christian Party (Parkindo)
SK Trimurti	Q7387595	Q19730387	Indonesian Party
SK Trimurti	Q7387595	Q4203019	Indonesia Labor Party

Appendix G: Screenshot of education table

person	personid	edu_id	educ	country	pesantren_m	military_pol
Ali Sastroamidjojo	Q2669326	Q156598	Leiden University	Netherlands	F	F
AA Maramis	Q12682432	Q156598	Leiden University	Netherlands	F	F
Susanto Tirtoprodjo	Q6723714	Q156598	Leiden University	Netherlands	F	F
R. Syamsudin	Q17410975	Q156598	Leiden University	Netherlands	F	F
Teuku Muhammad Hasan	Q4259946	Q156598	Leiden University	Netherlands	F	F
Prijono	Q4378323	Q156598	Leiden University	Netherlands	F	F
Mohammad Ichsan	Q120203830	Q156598	Leiden University	Netherlands	F	F
Sartono	Q9333940	Q156598	Leiden University	Netherlands	F	F
Sudjono Djuned Pusponegoro	Q19723944	Q156598	Leiden University	Netherlands	F	F
Assaat	Q4808308	Q156598	Leiden University	Netherlands	F	F
Nasaruddin Umar	Q6966619	Q156598	Leiden University	Netherlands	F	F
Achmad Soebardjo	Q4250968	Q156598	Leiden University	Netherlands	F	F
Soepomo	Q4201183	Q156598	Leiden University	Netherlands	F	F
Abdulmadjid Djojoadiningrat	Q61119002	Q156598	Leiden University	Netherlands	F	F
Sutan Muhammad Zain	Q19749281	Q156598	Leiden University	Netherlands	F	F
Nazir Datuk Pamoentjak	Q12500218	Q156598	Leiden University	Netherlands	F	F
Ferdinand Lumban Tobing	Q11188587	Q1934491	STOVIA (School for the Training of Native Doctors)	Indonesia (Dutch)	F	F
J. Leimena	Q2642883	Q1934491	STOVIA (School for the Training of Native Doctors)	Indonesia (Dutch)	F	F
Kasman Singodimedjo	Q16187634	Q1934491	STOVIA (School for the Training of Native Doctors)	Indonesia (Dutch)	F	F
Bahder Djohan	Q4842524	Q1934491	STOVIA (School for the Training of Native Doctors)	Indonesia (Dutch)	F	F
Lanjumin Dt. Tumangguang	Q19728278	Q1934491	STOVIA (School for the Training of Native Doctors)	Indonesia (Dutch)	F	F
Abdul Moeloek	Q19737572	Q1934491	STOVIA (School for the Training of Native Doctors)	Indonesia (Dutch)	F	F
Hazairin	Q11168814	Q19752649	Rechtshoogeschool te Batavia (Batavia Law School)	Indonesia (Dutch)	F	F
Mohammad Yamin	Q3503054	Q19752649	Rechtshoogeschool te Batavia (Batavia Law School)	Indonesia (Dutch)	F	F
Moeljatno	Q6890147	Q19752649	Rechtshoogeschool te Batavia (Batavia Law School)	Indonesia (Dutch)	F	F
Ide Anak Agung Gde Agung	Q981536	Q19752649	Rechtshoogeschool te Batavia (Batavia Law School)	Indonesia (Dutch)	F	F
Soemarno (ekonom)	Q19752816	Q19752649	Rechtshoogeschool te Batavia (Batavia Law School)	Indonesia (Dutch)	F	F
Joko Widodo	Q3318231	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Oemar Seno Adji	Q12501175	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
P.C. Harjasudirdja	Q25463109	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Rusiah Sardjono	Q12509201	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Ben Mang Reng Say	Q4886112	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Bambang Kesowo	Q12473890	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Boediono	Q76362	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Manuel Kaisiepo	Q12496803	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Mohamad Prakosa	Q12498747	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Soenarno	Q12515766	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Ali Ghufron Mukti	Q12471480	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Andi Alfian Mallarangeng	Q11093005	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Djoko Kirmanto	Q11182889	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Fadel Muhammad	Q5429207	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Gusti Muhammad Hatta	Q12484988	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Muhaimin Iskandar	Q12499327	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Patrialis Akbar	Q5247076	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Roy Suryo	Q9019254	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Wiendu Nuryanti	Q12524514	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Abdul Rahman Saleh (jaksa)	Q4665652	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Bambang Sudibyo	Q11120678	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F
Siti Fadilah	Q3118591	Q1145992	Gadjah Mada University (UGM)	Indonesia	F	F

Appendix H: Screenshot of family table

person	personid	relative	rel_id	status	distance_birth	distance_nuclear	family
I Gusti Ngurah Jaya Negara	Q102984055	I Gusti Ayu Bintang Darmawati	Q72117992	younger sibling	2	1	T
Panusunan Pasaribu	Q104597648	Dolly Pasaribu	Q105764924	child	1	1	T
Panusunan Pasaribu	Q104597648	Syahrul M. Pasaribu	Q17411516	younger sibling	2	1	T
Panusunan Pasaribu	Q104597648	Gus Irawan Pasaribu	Q31174037	younger sibling	2	1	T
Panusunan Pasaribu	Q104597648	Bomer Pasaribu	Q12476886	older sibling	2	1	T
Soeweno	Q105342433	Endang Kusuma Inten Soeweno	Q56399199	spouse	1	1	T
Soekardi	Q105427014	Theo L. Sambuaga	Q17411225	child-in-law	2	2	T
Dolly Putra Parlindungan Pasaribu	Q105764924	Panusunan Pasaribu	Q104597648	father	1	1	T
Dolly Putra Parlindungan Pasaribu	Q105764924	Bomer Pasaribu	Q12476886	uncle	2	2	T
Dolly Putra Parlindungan Pasaribu	Q105764924	Syahrul M. Pasaribu	Q17411516	uncle	2	2	T
Dolly Putra Parlindungan Pasaribu	Q105764924	Gus Irawan Pasaribu	Q31174037	uncle	2	2	T
Nazyra C. Noer	Q106450297	Arifin C. Noer	Q4790571	father	1	1	T
Nazyra C. Noer	Q106450297	Jajang C. Noer	Q6124331	mother	1	1	T
Paulus Pandjaitan	Q106462747	Faye Simanjuntak	Q72456512	nephew	2	2	T
Paulus Pandjaitan	Q106462747	Maruli Simanjuntak	Q19728927	sibling-in-law	2	2	T
Nina Agustina	Q106878555	Da'i Bachtiar	Q5207158	father	1	1	T
Zeke Khaseli	Q106977528	Ladya Cheryl	Q16172449	spouse	1	1	T
Zeke Khaseli	Q106977528	Agum Gumelar	Q4694586	father	1	1	T
Zeke Khaseli	Q106977528	Linda Amalia Sari	Q12494918	mother	1	1	T
Ridho Rahmadi	Q107061358	Tasniem Fauzia Rais	P44	spouse	1	1	T
Ridho Rahmadi	Q107061358	Amien Rais	Q2202270	parent-in-law	2	2	T
Ridho Rahmadi	Q107061358	Ahmad Hanafi Rais	Q16164615	sibling-in-law	2	2	T
Ridho Rahmadi	Q107061358	Ahmad Mumtaz Rais	Q16164694	sibling-in-law	2	2	T
Ridho Rahmadi	Q107061358	Rangga Almahendra	Q61118830	sibling-in-law	2	2	T
Ridho Rahmadi	Q107061358	Abdul Rozaq Rais	Q16162485	uncle-in-law	3	2	T
Ridho Rahmadi	Q107061358	Hanum Salsabiela Rais	Q7347498	sibling-in-law	2	2	T
Hanindhito Himawan Pramana	Q107564265	Pramono Anung	Q12506342	father	1	1	T
Abdul Latif Amin Imron	Q107994307	Fuad Amin Imron	Q19943268	older sibling	2	1	T
Tuti Sutiawati	Q108483499	Firman Santyabudi	Q16177287	child	1	1	T
Tuti Sutiawati	Q108483499	Kunto Arief Wibowo	Q61119148	child	1	1	T
Tuti Sutiawati	Q108483499	Try Sutrisno	Q76333	spouse	1	1	T
Rommy Sulastyo	Q108522104	Raissa Anggiani	Q109429993	child	1	1	T
Rommy Sulastyo	Q108522104	Annisa Trihapsari	Q10956050	younger sibling	2	1	T
Syahrul Yasin Limpo	Q10860301	Ichsan Yasin Limpo	Q16182111	younger sibling	2	1	T
Syahrul Yasin Limpo	Q10860301	Dewie Yasin Limpo	P29	older sibling	2	1	T
Syahrul Yasin Limpo	Q10860301	Adnan Purichta Ichsan	Q65212430	nephew	2	2	T
Syahrul Yasin Limpo	Q10860301	M. Yasin Limpo	Q26818531	father	1	1	T
Afriansyah Noor	Q108889808	Sidi Tando	Q14917468	grandparent	2	2	T
Raissa Anggiani	Q109429993	Rommy Sulastyo	Q108522104	father	1	1	T
Raissa Anggiani	Q109429993	Annisa Trihapsari	Q10956050	aunt	2	2	T
Indrata Nur Bayuaji	Q109430785	Susilo Bambang Yudhoyono	Q57405	uncle	2	2	T
Andi Sinjaya Ghalib	Q109432685	Andi Muhammad Ghalib	Q12472022	father	1	1	T
Abdul Syukur	Q109433339	Wahidin Halim	Q17411648	older sibling	2	1	T
Abdul Syukur	Q109433339	Hassan Wirajuda	Q183581	older sibling	2	1	T
Depriwanto Sitohang	Q109433358	Johnny Sitohang	Q17410631	father	1	1	T
Depriwanto Sitohang	Q109433358	Jonathan Ompu Tording Sitohang	Q112933079	grandparent	2	2	T
Ipuk Fiestiandani	Q109438116	Abdullah Azwar Anas	Q16162595	spouse	1	1	T
Danukromo	Q109439782	Ali Sastroamidjojo	Q2669326	great-grandparent	3	3	T
Danukromo	Q109439782	Danurejo I	P54	grandparent-in-law	3	3	T
Danukromo	Q109439782	Basyeiban	Q12474917	ancestor			T
Danukromo	Q109439782	Hamengkubuwono II	Q2509252	parent-in-law	2	2	T

Appendix I: Details of error checking for "multi-row" tables

I.1. Party table

We can see that most of the defective rows in this table are comprised of those that were not identified by our checking function. Some of these were those that referenced independents. Since Wikipedia does have pages for a wide variety of things that contain the word independent like "Independent politician" or "Independent events," what should have been a single category "Independents" ended up having many different unique identifiers. This is because my code generated unique IDs by looking up values in Wikipedia and returning the Wikidata ID associated with the page that was obtained. The solution here was simply to link all independents with a new, single custom variable for all independents with its own unique ID. Additionally, when joining the party table to the alias table I used for translating party names [3], some of the existing parties did not find matches in the alias table. When this happened, it was always because the Wikidata IDs in the original table referred not to specific party pages but disambiguation pages instead. For example, a row with the party name "PDI" should link to the page for *Partai Demokrasi Indonesia*¸ but in this case the link was to the disambiguation page for PDI. As for those rows that were marked by my checking function, these did usually reflect idiosyncratic structures that my regular expression-based code could not parse well.

Fortunately, it should be relatively straightforward to tweak our code to capture and account for these errors. It should be more than doable for instance, to build in processes to automatically link independents to the single, unique identifier I created. Most of the rows marked by the checking function also have certain recurring patterns that can be added to the

[3] For this, I created an alias table where each row represented every unique party that has ever existed in Indonesia. Fortunately, all of these parties had unique Wikipedia pages.

existing parsing code. Some will remain unaccounted for, but this appears to be a very minor proportion of the table. As for disambiguation pages, it should again be easy to write a function that can identify whether an ID in our database is referring to a disambiguation page. Since in practice only a few unique party abbreviations such as PDI end up being linked to disambiguation pages, it shouldn't be too much of a hassle either to manually create certain rules for these few instances. Overall then, it looks like we can reasonably cut down the amount of errors for the party table by a significant amount with a few refinements to our code.

I.2. Education table

The way in which I calculated and corrected errors in the education table differs somewhat from the other tables outlined in figure 5. In reality, my checking function marked 33.5% of this table's rows as suspicious. However, a lot of this was simply due to the fact that the checking function I devised for this table was perhaps overly stringent. One of the conditions I built into the function was to check whether the object entered in the education field had an establishment date, with the assumption that educational institutions should all have a recorded establishment date. The intention here was to ensure that all the entities linked to persons in this table would be actual educational institutions and not some bizarre entities unexpectedly captured by my parsing code. As it turns out however, the vast majority of the entities that returned valid Wikipedia pages in this table ended up being accurately linked to educational institutions. The only real errors that I had to correct, represented by the 2.4% of entries I outlined in figure 5, were hyperlinks to cities that would sometimes show up in individuals' Wikipedia pages. For instance, a hypothetical individual's education section might say "Leiden University, Netherlands" in his Wikipedia page. My code would capture Netherlands as well and return it as an entity in this table, since Netherlands has its own Wikipedia page as well. Again, we can easily build in functions that can detect locations in our code to deal with this. If not, these entries make up a sufficiently small portion of the table that manual cleaning should not be too time-consuming.

I.3. Positions table

A lot of the errors that showed up here were refreshingly straightforward. Most of the few rows that were marked by my checking function were those that did not return any Wikidata IDs for the listed "office," in some cases because these offices did not have Wikipedia pages, and in other cases because these entities were not really offices but some idiosyncratic values that my parsing function somehow extracted. These are exactly the type of irregularities that the checking function was designed to detect, and they make up a sufficiently small portion of the table that manually checking them should not be a problem.

A more interesting observation here was with regards to those defective rows that were not caught by the checking function. Almost all of these were position entities that were somehow linked to Wikidata IDs that for some reason or another corresponded to bizarre entities completely unrelated to our database. For instance, one of these IDs linked to a page listing the stars in the Hydrus constellation, another was linked to a page for a Japanese pop song. Perhaps some hidden link between these interesting concepts and our database does exist. Such a link escapes me however, and I proceeded by treating these entries as the strange and inevitable flukes that are part and parcel of scraping and deleting them. Again, these comprise a tiny portion of the entire table and can be safely deleted manually.

I.4. Family table

The family table contains by far the largest number of suspicious entries (figure 5). It should be noted that our checking function originally only marked 14.8% of the rows in our table. The reason why I expanded this significantly to cover 40.1% of the table is because often times, the existence of suspicious parsing structures in a single relation may signal odd parsing structures for all the relationships that specific person has. As such, I set a condition where if a certain person has one suspicious relation, all his relations should be marked as suspicious as well. To illustrate, the original checking function only marked one of Kaesang Pangarep's many relations as odd: namely, the row specifying that Kaesang had a spouse called "Blog video." The checking function was correct to mark this as suspicious. However, it failed to notice a separate curio: namely, that Kaesang's uncle, Anwar Usman, was incorrectly marked as his spouse in another row. That "Blog video" was able to make its way into the family dataset is itself indicative that something odd might be happening with Kaesang's infobox structure.

Nonetheless, it is also true that the checking function I ended up implementing for this version of the database may have been overkill. In many cases for instance, the entries marked by the checking function did not necessarily indicate strange infobox structures. Many of these were simply the result of careless parsing on my end. One of the checking filters I used was to check whether the obtained relative to a person had a birth date in his/her Wikipedia page, again with the assumption that only persons had such information. Unfortunately, in my code I only specified "birth_date" as a potential name of the field. In reality, there were multiple alternatives to specifying this field such as "dateofbirth" or "birthdate." Many of the relations marked by my checking function had birth date fields that I had failed to specify. This is easily remediable, and also does not warrant an expansion of checks beyond the singular marked relation since this error does not necessarily reflect a broader idiosyncrasy within the person’s entire infobox. Additionally, many of the marked entries simply did not have Wikipedia pages. Again, it will not be necessary to assume some strange infobox structure for such individuals’ entire family relations. By refining the parsing code to reflect this, we can create a more targeted checking function and reduce the number of unnecessary manual checks.

Finally, for several relations my parsing function only specifies a general relationship status such as “parent” (as opposed to “mother” and “father”) or “relative” (which could really mean anything). Again, the substantial inconsistency in the way infobox data pertaining to family relationships is structured is the key culprit here. Specifying overly strict parsing functions risks extracting false information. This is exacerbated by the fact that unlike persons, family relations generally do not have hyperlinks to Wikipedia articles that we can use as a sanity check to see whether the extracted relation titles are accurate or not. As such, my compromise was to simply specify a relation in very general terms when the structure doesn’t appear to be clear. In these cases, we will have to manually check each relation status and change it to its more accurate, precise form. At the moment, there appear to be no workarounds to this unfortunately, and some degree of manual cleaning for this specific field appears to be inevitable.

I.5. Summary

In short, these are the assumptions I made in calculating the hypothetical amount of cleaning necessary once we’ve made the tweaks specified above to our parsing code:

- **Party:** I assume that all of the defects related to disambiguation pages and independents can be solved, leaving only the idiosyncrasies captured by the original checking function.

- **Education:** If we are able to build a parsing function that handles locations, we are only left with very few idiosyncrasies captured by the checking function.
- **Positions:** No change from the previous value, most of the manual checking being done consists of unpredictable idiosyncrasies already.
- **Family:** Most of the reduction comes from narrowing the checking function down in the way outlined above. Manually changing vague relationship titles such as “parent” and “relative” will still have to be done fully.

Appendix J: Note on Code Appendix

All the code used in the course of this study will be provided in separately attached HTML renderings of Jupyter Notebooks. There will be three separate code appendices attached:

- **Code Appendix A:** The main code used for generating the database
- **Code Appendix B:** The code used for generating the original structures of our alias-translation tables, and joining them to our main tables
- **Code Appendix C:** The code used for the analysis conducted in sections 4.2 and 4.3 above