Social Networks and the 
Political Salience of Ethnicity

Nicholas Eubank*

August 17, 2018

Forthcoming, Quarterly Journal of Political Science

Abstract

Ethnic politics scholars are increasingly convinced that (a) the political salience of ethnicity and (b) the correlation between ethno-linguistic fractionalization (ELF) and poor development are driven by the dense social networks shared by co-ethnics. By this argument, social networks allow ethnic parties to leverage inbuilt networks to share information and support collective action, while ethnically fragmented communities struggle to hold politicians accountable. This paper provides the first comprehensive empirical test of the assumption underlying this argument. Using seven months of telecommunications data from 9 million mobile subscribers in Zambia – which includes records of almost 2 billion calls and SMS messages – to measure social networks across an entire country, this paper finds that electoral constituencies with high ELF also have more fragmented social networks, especially in rural areas. It also finds other potential cleavages that have not achieved political salience (namely religious identity and employment sector) are not correlated with network fragmentation, consistent with the idea that ethnicity achieves salience because it offers an organizational advantage not offered by other cleavages. Finally, it also finds that both voter knowledge and public goods are negatively correlated with network fragmentation, consistent with the network-proxy hypothesis.

*Center for the Study of Democratic Institutions, Vanderbilt University. nick@nickeubank.com. This project would not have been possible without the generosity of numerous parties, including Real Impact Analytics, Real Impact’s Chief Data Scientists Gautier Kring, and its CEO Sébastien Deletaille. The author is also indebted to Ramya Parthasarathy, Ernest Mundia, Neil Malhotra, Katherine Casey, Jim Fearon, Kate Baldwin, the Zambian Central Statistics Organization, Adriane Fresh, Vincent Traag, Jeremy Weinstein, Jonathan Rodden, Viji Jagan-nathan and the Stanford GSB Center for Global Business and the Economy, Donovan Neale-May, Jure Leskovec, Kate Baldwin, Romain Wacziarg, Tamás Nepusz, James Moody, and participants in the Stanford GSB Organizational Behavior seminar, the 2015 WGAPE Meeting at Brown University, and Duke Geospatial Data, Governance, and the Future of Development Aid conference, and numerous anonymous referees. Financial support provided by a Stanford Vice-Provost for Graduate Education Diversity Dissertation Research Opportunity grant, a Stanford GSB PhD Research grant, a Stanford Center for Philanthropy and Civil Society (PACS) Fellowship, and a Dolores Zohrab Liebmann Fellowship. Additional acknowledgements in Appendix V.
Research on the politics of ethnicity is dominated by efforts to explain two prominent empirical regularities. First, across a broad range of regions and institutional contexts, ethnicity consistently achieves a high degree of political salience (e.g. Chandra (2007); Posner (2005); Basedau et al. (2011); Keefer (2010); Laitin (1998); Horowitz (1985)). This is true not only in terms of political rhetoric, but also political party organization and voting behavior. And second, ethnically fragmented communities tend to experience worse development outcomes and low investment in public goods (e.g. Easterly and Levine (1997); Alesina, Baqir and Easterly (1999); Alesina and La Ferrara (2005); Miguel and Gugerty (2005); Khwaja (2009); Habyarimana et al. (2007)).

In search of explanations for these regularities, scholars increasingly suggest that the salience of ethnicity is not driven by ethnicity per se, but rather social network structures that are assumed to be closely related to ethnicity. For example, Miguel and Gugerty (2005, p. 2330), argue “that social sanctions and coordination are possible within [ethnic] groups due to the dense networks of information and mutual reciprocity that exist in groups but are not possible across groups.” Similarly, Habyarimana et al. (2009) argue that if co-ethnics are more well-connected in terms of their social networks, then co-ethnics may be more “findable” and thus subject to social sanctioning. And in the literature on ethnic violence, Fearon and Laitin (1996, p. 719) assert that ethnic groups are “often characterized by relatively dense social networks,” and that “across groups […] social networks are less developed” so that “it is more difficult to get information on a potential trading or social partner from ‘across the tracks.’ ”

The political networks literature – which documents ways in which network structure shapes capacity for political action – provides additional support for this explanation. Well-connected social networks may help communities overcome collective action problems associated with getting constituents to political rallies or protests (Jackson, Rodriguez-Barraquer and Tan 2012; Habyarimana et al. 2009), and to turnout on election day (Rojo and Wibbels 2014; Cox 2014).

1Note that this literature does not tend to take a strong position on why ethnicity and network structure are closely related. These explanations are consistent with the possibility that values or other factors related to ethnicity shape network structure, which in turn drives political salience.
Well-connected networks may also support information diffusion (Larson 2017b; Larson and Lewis 2017), making it easier for politicians to learn what citizens want and develop appealing policy platforms, and for citizens to monitor elected officials and hold them accountable. Well-connected networks may also help voters coordinate their support around a candidate, and avoid vote-splitting.

Taken together, the link between ethnicity and network structure, on the one hand, and network structure and capacity for political activity on the other, have the potential to explain both empirical regularities noted above – the political salience of ethnicity and the tendency for ethnically fragmented communities to experience worse development outcomes. If ethnic groups share well-connected networks that support political action, for example, then political parties that organize along ethnic lines will inherit a free organizational advantage over non-ethnic parties. This would make ethnic parties more competitive than non-ethnic parties. And if ethnically fragmented communities lack networks that support collective political behaviors, they would also have difficulty holding politicians accountable, leading to shirking, underinvestment in public goods, and poor development outcomes.

But while this network-proxy hypothesis has generated substantial excitement in the ethnic politics literature in recent years (e.g. Habyarimana et al. (2009, 2007); Miguel and Gugerty (2005); Larson (2017a); Rojo and Wibbels (2014); Dionne (2015)), the motivating assumption of this argument – that ethnically fragmented groups have more fragmented social networks – has never been systematically tested. Instead, most existing work (like Miguel and Gugerty (2005) and Fearon and Laitin (1996) quoted above) simply assumes this relationship to exist. And while there are reasons to believe this may be the case, much of the evidence cited in support of this assumption – like evidence of ethnic homophily – is not sufficient to substantiate it (as discussed in Section 2). As a result, this growing literature rests on a dangerously untested foundation.

Though Larson and Lewis (2017) offers a theoretical nuance to this argument, along with preliminary empirical evidence suggested where links differ qualitatively, density of ties may not be monotonically associated with faster information diffusion.
This paper fills that gap by providing the first comprehensive test of this literature’s underlying assumption: that ethnically fragmented communities have more fragmented social networks. I use seven months of detailed mobile telecommunications data to measure the structure of social networks in each of Zambia’s 150 National Assembly electoral districts. In particular, I use this data – which includes records of almost 2 billion calls and SMS (text) messages – to construct a measure of network fragmentation for each district. By pairing this measure with geo-coded census data on the ethnic composition of districts, this paper systematically examines the relationship between ethnic fractionalization and network fragmentation across the universe of electoral constituencies in Zambia.

In doing so, this analysis joins three other studies aiming to test this assumption. First, Larson and Lewis (2017) map social networks in two villages in Uganda and find networks are denser in an ethnically heterogeneous village, though they also find information diffuses more effectively in a homogeneous village. Dionne (2015) maps networks in four villages of Malawi and finds cross-ethnic ties to be as common as intra-ethnic ties. And finally Habiyarimana et al. (2009) finds that lab subjects find random strangers more quickly if they are co-ethnics, suggesting co-ethnics have more well-connected networks. Unlike past studies, however, this analysis is not limited by the logistical constraint of studying a relatively small number of geographically-confined networks, or to measuring small-scale network properties. Instead, using cell-phone meta-data this analysis is able to measure the social networks of an entire country, and study variation in network structure at the most politically-relevant scale for theories of party organization and political accountability – at the scale of electoral constituencies.

This analysis finds that ethno-linguistic fractionalization (ELF) and network fragmentation are generally positively correlated, as predicted by the network-proxy hypothesis. This result is especially strong in rural constituencies, suggesting that urbanization may diminish the salience of ethnicity. It also finds that network-fragmentation is not correlated with other social divisions that have failed achieved political salience in Zambia, like religious and economic divisions,
consistent with the idea that ethnicity offers organizational advantages to voters and party leaders that other organizing cleavages do not. And finally, this analysis also examines whether network fragmentation is associated with lower voter knowledge and public goods provision, as the network-proxy hypothesis suggests. It is, albeit not as strongly as network fragmentation is associated with ELF. Finally, as detailed in Section 7 the interrelationship of public goods, voter knowledge, ELF and network fragmentation is also consistent with the network-proxy hypothesis.

Throughout this analysis, every effort is made to provide an authentic test of the network-proxy hypothesis as put forth by past scholars. The measure of network fragmentation used in this analysis, for example, is constructed to measure the properties of networks past scholars have suggested should vary with ethnic composition as closely as possible. Similarly, ethnic fractionalization is measured using the canonical measure of ethnic fractionalization used to establish the empirical regularities that the network-proxy hypothesis aims to rationalize – ELF. And finally, ELF is calculated with respect to a dimension of identity that has the key features of an ethnic cleavage that the network-proxy hypothesis aims to explain: it is based on ascriptive features that are not easily chosen or changed by individuals (thus meeting a standard definition of ethnicity (Fearon, 1999)) and it is a politically salient.

These results constitute the first direct validation of the assumption that ethnic fragmentation is associated with network fragmentation. This both affirms an untested yet critical assumption underlying numerous past studies – improving our confidence in those results – and also provides a firm foundation for future researchers that the motivating assumption of this literature is well-founded. Moreover, in showing that voter knowledge and public goods tend to be negatively correlated with network fragmentation, it provides further support for the network-proxy hypothesis. And finally, in finding that this relationship is particularly strong in rural communities, this research suggests a potential avenue for future research into the causes of the attenuation of the network-ethnicity relationship in urban communities, causes with may also have implications for efforts to diminish the political salience of ethnicity more broadly.
The validation of this assumption was by no means a forgone conclusion. Indeed, many other promising hypotheses aimed at explaining the salience of ethnicity that were premised on seemingly self-evident assumptions have struggled in the face of subsequent empirical interrogation. For example, the theory that ethnic coalitions reflect shared policy preferences (and therefore that bad outcomes in ethnically fragmented communities stem from policy disagreements (Alesina, Baqir and Easterly, 1999)) has been met with recent studies that show policy preferences do not appear to vary dramatically across ethnic groups (e.g. Desmet, Ortuno-Ortín and Wacziarg (2015); Habyarimana et al. (2009). See Lieberman and McClendon (2012) for a dissenting view.), and ethnic parties do not appear to be structured to maximize policy influence (Keefer, 2010). The idea that people simply care more about co-ethnics than non-co-ethnics (Charness and Rabin, 2002; Chen and Li, 2009) and so under-invest in public goods that also benefit non-co-ethnics (Ejdemyr, Kramon and Robinson, 2015) also does not square with laboratory results that show that participants in anonymous altruism games seem just as generous to co-ethnics as non-coethnics (Habyarimana et al., 2009; Berge et al., 2015; Dionne, 2015). And finally, the idea that ethnicity is strategically valuable because the recognizability and immutability of ethnicity allow minimal winning coalitions to ensure new-comers do not sneak in to claim benefits (Fearon, 1999) has been challenged by evidence that in many African contexts ethnicity is not that easy to identify in strangers (Habyarimana et al., 2009; Casey, 2016; Harris and Findley, 2012).

1 Ethnicity in Zambian Politics

Ethnicity has always been a core organizing concept of political debate in Zambia. Today, the political salience of ethnicity is evident not only in aggregate voting patterns (Posner, 2005; Erdmann, 2007) but also in individual-level survey data. Using data from a 1996 post-election survey, Posner and Simon (2002) conclude that while economic conditions do matter, “ethnicity and urban/rural location explain the lion’s share of the variance in patterns of support for the
incumbent regime” (Posner and Simon, 2002, p.329). Similarly, using data from the 2004 wave of Afrobarometer, Basedau et al. (2011) examine the relationship between ethnicity and vote choice. They conclude that while ethnicity is not a strong predictor of political preferences in all African countries, Zambia is subject to “comparatively strong or medium to strong ethnicization of party politics,” (Basedau et al., 2011, p.467) based on their ability to predict vote choice using ethnicity. Indeed, they note that ethnicity was a significant predictor of voting preferences for all political parties they examined (MMD, UPND, UNIP, and PF). And finally, Erdmann (2007) reaches similar conclusions based on data from focus groups and field surveys: “ethnicity can still be viewed as the major factor explaining party affiliation in Zambia, and to a lesser degree voter alignment.” (p. 30). The same author reports that in focus groups, “most of the participants could not detect any programme or policy difference between the parties. And the few who said they could detect differences were, when directly challenged, almost all unable [sic] to name any difference” (Erdmann, 2007, p.23).

Despite agreement on the political salience of ethnicity in Zambia, the question of why ethnicity is politically salient remains largely unanswered. As Laitin (1998, p. 248) has argued, “ethnic entrepreneurs cannot create ethnic solidarities from nothing. They must, if they are to succeed, be attuned to the micro incentives that real people face.” So what micro-incentives are at play in Zambia that make ethnicity so preferable to a different cleavage? Class differences, for example, are widely noted in Zambia and have clear policy implications, and yet have never achieved political salience (Posner, 2005; Chikulo, 1988).

---

Posner (2005) provides a partial answer to this question in the Zambian context, arguing that strategic political considerations drive the specific dimensions of identity that become politically salient. Yet this explanation begins from the assumption that citizens will mobilize along some dimension of identity. The one exception to this oversight comes in a discussion of why Zambians do not organize along class lines. Here, Posner (2005, p. 86) argues “[t]he answer is that, by and large, class identities are not sufficiently deeply felt for them to play this role.” This implies that it is the emotional salience of ethnicity that leads to its viability as a political cleavage, implicitly supporting the social-psychological mechanism discussed below.
Explaining the Salience of Ethnicity

The network-proxy hypothesis provides one potential explanation. If ethnicity is closely tied to network structure, then ethnically aligned constituencies may be better able to overcome the collective action problem, coordinate activities, and share information.

Consider, first, the collective action problem. All supporters (or opponents) of a candidate or cause benefit from rallies and protests whether they participate or not, so many individuals may prefer to free-ride on the attendance of others, even if they support the cause. Social pressure on non-contributing group members can solve this problem (Ostrom 1990; Jackson, Rodríguez-Barraquer and Tan 2012; and Wolitzky 2012) have shown that the ability to apply social pressure is a function of network connectedness.

Social networks may also help to overcome coordination problems, like rallying voters around a specific candidate (Siegel 2009). When more than two candidates from a given community stand for election, avoiding vote splitting is critical to political effectiveness. Moreover, well-connected social networks may also help overcome mundane logistical challenges like learning about voter preferences, coordinating on objectives, or scheduling and publicizing political events.

In addition, social networks may help facilitate information diffusion, which has implications for electoral accountability. Theoretical and empirical research has consistently found that citizens are only able to hold politicians accountable when they know what those politicians are doing (e.g. Besley 2007; Reinikka and Svensson 2004; Ferraz and Finan 2008, among many others). Citizens cannot reward good behavior and punish poor behavior if they cannot observe politician behavior in the first place. Moreover, poor information diffusion also makes the application of social pressure more difficult (Fearon and Laitin 1996; Larson 2017b).

These factors help to explain why political entrepreneurs may choose to organize along ethnic lines as opposed to along other cleavages: ethnic coalitions come with inbuilt networks that enhance political effectiveness. Moreover, this advantage is likely to be especially acute in de-

---

4 Studies of voting also suggest social pressure increases turnout (Gerber, Green and Larimer 2008).
veloping countries, where political parties tend to be poorly institutionalized and underfunded. Parties organized along other dimensions can develop their own machinery, of course, but doing so diverts scarce resources from other political activities, undermining competitiveness, and making this option less desirable.

These factors also help to explain why citizens may be willing to mobilize along ethnic lines. By joining ethnic coalitions, voters access effective political organizations that can better share information and sanction politicians (through protest or coordination around alternative candidates), likely leading to better governance.

And finally, these factors may also explain why ethnically fragmented communities have difficulty holding politicians accountable, resulting in lower investment in public goods and poor development outcomes.

These insights help to explain why, if ethnicity is closely tied to network structure, it makes sense for ethnicity to achieve such political salience in various settings. Despite the promise of this explanation, however, the network-proxy hypothesis is arguably the least-tested explanation for ethnic salience – a short-coming remedied here.

2 Measuring Social Network Fragmentation

Existing work has argued that co-ethnics tend to be well-connected to one another, and that they tend to have more connections with one another than with non-co-ethnics (Fearon and Laitin 1996; Miguel and Gugerty 2005; Habyarimana et al. 2009). With that in mind, this section develops a measure of network fragmentation that reflects this characterization as closely as possible in three steps.

First, I construct a social network from cell-phone meta-data. Each cell-phone subscriber forms a node or vertex in this network, and connections are placed between individuals who call or text one another. This process is discussed in more detail in Section 3. Second, this social network of all Zambian cell-phone subscribers is partitioned into groups
so that (a) the members of each group are as well-connected to one another as possible, and (b) there are as few connections running between groups as possible. If current characterizations of ethnic networks are accurate, then these groups – called network communities – should roughly reflect ethnic divisions.

This partitioning is based only on patterns of interaction. The network will be partitioned along whatever lines are most clearly defined in the network. If, as network-proxy theorists have argued, ethnicity is the dominant determinant of network divisions (and that it is for this reason that ethnicity is more likely to achieve political salience than other potential cleavages), then these network communities should mirror ethnic divisions. However, if other divisions (like class) play a larger role in shaping patterns of interaction, then these partitions will not reflect ethnic divisions.

In the third and final step, I compute a measure of network fragmentation at the level of electoral constituencies. More specifically, I calculate network fragmentation as one minus the Herfandahl index of network community assignments for residents of an electoral constituency. This measure is perfectly analogous to ELF, which is computed as one minus the Herfindahl index of individuals’ ethnicities. Indeed, if the network has been partitioned into network communities that parallel ethnic divisions, then these two measures should be roughly equivalent and thus highly correlated in subsequent analyses.

Note that in the second step, the partitioning algorithm is applied to the global network of Zambia; that is, it does not take into account information about the spatial distribution of users. Only after users are partitioned is information about user residency used. I employ this global partition strategy rather than the alternative — a constituency partition strategy, in which sub-networks consisting only of the residents of each electoral constituency are created and partitioned — for several reasons. First, it best utilizes all of the available data. A discussed in more detail below, while this analysis makes use of network meta-data for all Zambian cell-phone subscribers, only subscribers of the Partner Telecom (PT) who have provided the cellphone meta-data used in this analysis can be geo-referenced. Thus as as non-PT subscribers could not
be assigned to a constituency, a constituency partition strategy that first subsets on residents of a constituency would require dropping a significant portion of the available network data. This would not only reduce the statistical power of estimates of network structure, but also likely lead to biased estimates, as most aspects of network topology are not preserved under sampling, even when that sampling is i.i.d. (Kolaczyk [2005] Leskovec and Faloutsos, 2006). A global partition, by contrast, comes much closer (as close as possible in this context) to capturing the full network.

Second, the global partition strategy allows me to capture cross-constituency and cross-village ties which can provide important information about the strength of social ties. For example, suppose $i$ and $j$ live in Constituency 1, and are both friends with $a$, $b$, and $c$, who live in Constituency 2. A global partitioning strategy allows for these out-of-constituency ties to increase the likelihood that $i$ and $j$ will be assigned to the same network community. A constituency partitioning strategy, by contrast, would ignore mutual connections to $a$, $b$ and $c$.

And finally, the global partition strategy allows me to examine the spatial distribution of members of each network community as a test of measure validity, as detailed in Section 4.

It is important to emphasize that the measure of network fragmentation created through these three steps is qualitatively different from measures used in most studies of ethnicity and networks. Due primarily to data constraints, most studies rely on local network measures (like homophily) that only consider individuals and their immediate connections. The network fragmentation measure used in this analysis, by contrast, takes into account not only local network features, but also meso-scale network structure: the structure of the network at the level of medium to large groups (Jackson, Rogers and Zenou 2017). As most network-proxy theories are about group-level dynamics, this is critical to properly testing the network-proxy hypothesis.

**Assignment to Network Communities**

Individual subscribers are partitioned into network communities using a Constant Potts Model (CPM). As described in detail in Appendix D, the CPM assigns each subscriber to one
and only one network community so as to maximize the density of links within each network community and minimize the number of links running between communities.

While most technical aspects of the CPM can be relegated to appendices, there is one parameter in the model that it is critical to discuss. When assigning individuals to network communities, the CPM attempts to meet two conflicting criteria: (a) maximize the density of links inside each community, and (b) minimize the number of links between groups. The more the model attempts to maximize (a), the more it will assign individuals to small, internally-densely connected groups; the more the model attempts to minimize (b), the more it will assign individuals to large, internally-loosely connected groups. As such, the model requires specification of a resolution parameter (commonly denoted $\gamma$) that determines the trade-off between these two goals.

Traditionally, specification of the resolution parameter is treated as a nuisance and an estimation challenge (referred to as the problem of scale). It is addressed by choosing the value of $\gamma$ that gives rise to the most (myopically) statistically significant community structure, meaning the community structure which is least likely to emerge in a network where edges are randomly distributed between nodes.\(^5\)

While convenient, this approach is deeply problematic for social scientists. The resolution parameter reflects a deep and substantively-important truth about communities: communities can exist on many different scales, and the question of which scale is most relevant depends on the question being asked. In the study of intra-village politics, for example, we may be interested in the structure of the small communities that form within villages; in the study of

---

\(^5\)All community detection algorithms address the problem of scale. Modularity-optimizing community detection algorithms, for example, implicitly follow the strategy described above. Others, like InfoMap, use an analogous information-theoretic criterion.

\(^6\)It is worth noting that this kind of problem of scale is not unique to network analysis, and these two approaches employed here are portable to other contexts. An analogous problem exists in many inductive algorithms, including any clustering algorithm in which individual observations are inductively grouped into buckets based on similarity. In text-as-data topic models, for example, the number of topics allowed is a near perfect analogue. Similarly, the results of spatial clustering analyses depend on the scale at which clustering is measured – clustering of voters is very different within cities than at the level of states, for example. As such, discussion of how best to address the problem of scale is of broad and increasing relevance as political science adopts more and more tools from statistics and computer science.

12
national politics, by contrast, we are likely to be more interested in community structure at the level of large regions.

This paper departs from reliance on this myopic statistical criterion (heretofore referred to as the atheoretic criterion) in two ways. First, network communities are calculated for values of the resolution parameter $\gamma$ across the parameter space (from small, internally dense communities to large, looser communities), and results are presented for all values. This not only provides a sense of robustness of results, but also a novel opportunity to gain insights into the scale at which network structure appears to matter for political outcomes, turning a nuisance parameter into an opportunity for learning.

Second, as discussed in more detail in Section 7, this analysis also uses variation in scale to further test the network-proxy hypothesis. If ethnicity achieves political salience because ethnic fractionalization is correlated with network fragmentation, and network fragmentation impedes the ability of communities to effectively engage in political activity, then not only should ELF and network fragmentation be correlated, but the scale (value of $\gamma$) at which they are correlated should also be the scale at which network fragmentation is negatively correlated with political outcomes like public goods and voter knowledge.

**Geo-referencing Subscribers**

The output of the CPM is an assignment of each individual subscriber to a network community. In order to convert these individual-level assignments into information about network fragmentation in each electoral districts, it is first necessary to identify the electoral district in which each subscriber resides.

Geo-referencing of subscribers takes advantage of the fact that most calls are routed through

---

7To illustrate, consider a network of academics. Network communities that consist of members of a sub-discipline (like Americanists, Comparativists, and Normative Theorists in a political science department) may be the most clearly delimited groups in a myopic statistical sense (since we talk most to colleagues whose work is most related to our own). However, if we wish to use this network to predict voting behavior in a University-level debate over how funds should be allocated across departments, we know intra-departmental divisions are less likely to shape voting behavior than super-department network communities (social science departments, natural science departments).
the nearest antenna tower. The cell-phone meta-data used for this analysis identifies the antenna
tower that handles every call placed by a subscriber of the PT. Using this information, combined
with data on the GPS location of these antenna towers, it is possible to geo-reference cell-phone
subscribers to specific zones based on the assumption that a user’s home is closer to his or her
most used cell-tower than any other tower. This process is detailed in Appendix F, along with
information on several additional refinements—such as restricting attention to calls made before
8am and after 6pm (when they are most likely at home and not at work) and using information
on the second-most-used cell-phone tower.

**Measuring District Fragmentation**

The previous two sections detail how each individual cell-phone subscriber is inductively as-
signed to a network community, and how each individual cell-phone user is geo-referenced.
This section describes how these results are combined to create a measure of whether residents
of each electoral district are divided into many small network communities (high fragmenta-
tion), or into a small number of large network communities (low fragmentation). As previously
noted, this is computed in the same manner as ELF. So, if network communities are coincident
with ethnic groups, the network fragmentation measure computed here will be equivalent to
ELF.

This aggregation allows subsequent analyses at the level of the electoral constituency. Most
forms of political action—like rallying around an opposition candidate, or applying pressure
to an elected official—require participation of a substantial share of an elected official’s con-
stituents to be effective. As such, constituency-level measures of network structure constitute
the most appropriate unit of analysis for the study of ethnicity, network structure, public goods,
and voter knowledge.

As detailed in Appendix G, most public goods are under the control of Zambia’s national
government, and so the focus of this analysis is on the electoral districts (Constituencies) for
members of Zambia’s national legislature, the National Assembly. Zambia’s national govern-
ment has primary *de facto* authority over public goods in Zambia, including water, electricity, education, police, and more. As a result, the ability of citizens to influence their elected representatives in the National Assembly is likely to affect the quality of service they experience across nearly all government departments.

**Computation of District Fragmentation**

The district-level measure of network fragmentation is calculated in a manner analogous to ELF: for each electoral district \( d \) and community partition of the network into network communities \( c \in C \), network fragmentation \( (NF) \) is computed as one minus the Herfindahl index of communities:

\[
NF_d = 1 - \sum_{c \in C} \left( \frac{n_{c,d}}{n_d} \right)^2
\]

where \( n_d \) is the total number of cell-phone subscribers who reside in district \( d \), and \( n_{c,d} \) is the number of cell-phone subscribers in district \( d \) assigned to community \( c \). This generates a single measure of network fragmentation for each electoral district that can be interpreted as the probability that any two randomly selected residents of a district are in different network communities.

One nuance to this calculation is that available data can only geo-reference cell-phone subscribers to specific zones, not points, so adjustments must be made for situations where subscribers are assigned to zones that are not completely contained within an electoral district. As discussed in detail in Appendix M, this is accomplished by assuming users are distributed in proportion to the area of the assigned zone that falls within each district and in proportion to each district’s population density.

---

8For example, a crude geo-referencing would assign a user to the area in which the user’s most-used tower is the closest tower. Full details of the more detailed but analogous geo-referencing strategy used in this paper can be found in Appendix F.
3 Data

Cell-Phone Meta-Data

Network data for this analysis consists of records of all SMS (text) and voice transactions passing through the network of a major Zambian Telecommunications company from December 2011 to June 2012. This data – commonly referred to as Call Detail Records or CDRs – includes approximately 2 billion transactions. Each transaction includes type (voice call or SMS), anonymized identifiers for both caller and receiver, date and time of transaction, duration, and GPS coordinates of the antenna tower through which the call was placed. The anonymized identifiers for callers and receivers are stable codes that allow usage patterns to be tracked over time but cannot be used to identify individuals or their phone numbers.

This dataset comes from one of the three dominant cell-phone providers in Zambia. Because of the sensitivity of the data, the telecom providing the data has asked to be referred to only as the Partner Telecom (PT), and certain commercially-sensitive statistics (like the exact location or number of antenna towers) have been omitted or are presented in an intentionally imprecise manner. As of early 2012, it can be said that PT had a market share of 25% - 40% had between 500 and 700 antenna towers, and provided service in 148 of Zambia’s 150 Constituencies.

Two types of cell-phone users appear in the network data: PT subscribers and non-PT subscribers. Because the dataset is constructed by aggregating PT subscriber data, it includes (a) all transactions by PT subscribers, and (b) transactions between PT subscribers and non-PT subscribers. As a result, nearly all cell-phone users in Zambia appear in the data, including non-PT most subscribers. Not included in the data are (a) non-PT subscribers who never call PT subscribers, (b) calls that occur between pairs of non-PT subscribers, and (c) geo-spatial information on non-PT subscribers.  

---

9 The range of market shares for the top three telecoms in 2012.
10 The 2010 census puts Zambia’s population at 12.5 million individuals; and there are approximately 9 million users in the cell-phone meta-data.
11 Users are geo-referenced using information on the antenna towers to which they connect. Telecoms manage their own towers, so PT CDRs only include information on tower routing for PT subscribers.
These data make it possible to measure social networks in a fundamentally different way from past studies that collected data by asking individuals for the names of their closest friends in a geographically-bounded area (e.g. Banerjee et al. (2014); Rojo and Wibbels (2014); Fafchamps and Vicente (2013); Dionne (2015)) and provide several advantages. First, they provide exceptional breadth – they cover an entire country – making it possible to estimate the properties of networks in 150 national electoral districts (population ~100,000). This makes cross-sectional analyses of the relationship between network structure and outcomes of interest possible at the most politically-relevant level. Second, the data have exceptional depth, capturing even the weak social ties sociologists have found to be exceptionally important, but which are often censored or forgotten when individuals are asked to list their five or ten closest friends (Granovetter, 1973). Third, the data include network ties between villages and between villages and cities, providing a more complete picture of networks than geographically-bounded surveys. And finally, unlike self-reported data, these data measure actual communication patterns, reducing concern of under-reporting of inter-class or inter-ethnic relationships. This is not to say meta-data perfectly capture social networks – like all other methods of network mapping, it has limitations – but these shortcomings are balanced by the advantages meta-data provides, making it an important compliment to traditional survey mapping.

Because of these advantages, meta-data has become an increasingly popular tool for analyzing networks in recent years, especially among computer scientists and economists. It has been used to measure internal migration and mobility (Blumenstock, 2012; Wesolowski et al., 2012), estimate wealth and socio-economic status in hard-to-surveil contexts (Blumenstock, Cadamuro and On, 2015), gain insights into the topology of real-world networks (Onnela et al., 2007), examine the spatial distribution of network communities (Blondel, Krings and Thomas, 2010; Barthelemy, 2011), study social processes like the spread of knowledge through networks (Björkegren, 2015), and estimate the credit worthiness of unbanked populations (Björkegren and Grissen, 2015). To the best of this author’s knowledge, however, this is the first use of cell-phone meta-data for the study of political phenomena.
Table 1: Who do people call most?

<table>
<thead>
<tr>
<th>Share of Respondents</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Family members</td>
<td>51.2%</td>
</tr>
<tr>
<td>Friends</td>
<td>44.0%</td>
</tr>
<tr>
<td>Business clients</td>
<td>4.1%</td>
</tr>
<tr>
<td>Business suppliers</td>
<td>0.6%</td>
</tr>
<tr>
<td>Information services</td>
<td>0.2%</td>
</tr>
</tbody>
</table>

Source: Research ICT Africa (2008)

Cell Phone Use in Zambia

The rapid penetration of cell-phones in Zambia make them a useful tool for measuring social networks. According to a 2012 nationally-representative Afrobarometer survey of 1,200 adult citizens (Afrobarometer, 2012), households report an average of 0.79 phones per adult and 79.6% of households have at least one cell-phone. Even in rural communities, 67.0% of households report owning a mobile phone. Phones are also well-used. Among respondents in households with a phone, fully 83.0% report using the phone at least once a day. Moreover, data from Research ICT Africa (2008) also suggests that these phones are used extensively for social purposes, not just business, as shown in Table 1.

There is also little evidence of systematic differences in usage across ethnic groups. As shown in Table 2, phone ownership rates are relatively similar across ethnic groups. And while a limited sample size makes strong inferences difficult, as shown in Table 3, ethnic groups do not appear to sort into different carriers.

This high penetration, high level of use, and the fact phones are being used for social as well as business purposes suggests cell-phone meta-data provide a good source of data for studying social networks.

12 Most enumerators appear to have skipped the open-ended “Who is your primary carrier” question.
Table 2: Phone Ownership by Ethnicity

<table>
<thead>
<tr>
<th>Ethnic Group</th>
<th>Avg Num Phones Per Adult</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barotse</td>
<td>0.80</td>
</tr>
<tr>
<td>Bemba</td>
<td>0.80</td>
</tr>
<tr>
<td>NW</td>
<td>0.74</td>
</tr>
<tr>
<td>Nyanja</td>
<td>0.85</td>
</tr>
<tr>
<td>Tonga</td>
<td>0.72</td>
</tr>
<tr>
<td>Total</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Source: Afrobarometer (2012)

Table 3: Subscribers by Ethnicity

<table>
<thead>
<tr>
<th></th>
<th>CelTel</th>
<th>ZamTel</th>
<th>MTN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bemba</td>
<td>42.2%</td>
<td>25.0%</td>
<td>76.5%</td>
</tr>
<tr>
<td>Lozi</td>
<td>3.5%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>NW</td>
<td>3.5%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Nyanja</td>
<td>38.7%</td>
<td>71.4%</td>
<td>17.6%</td>
</tr>
<tr>
<td>Tonga</td>
<td>12.1%</td>
<td>3.6%</td>
<td>5.9%</td>
</tr>
<tr>
<td>No. Obs.</td>
<td>256</td>
<td>28</td>
<td>17</td>
</tr>
</tbody>
</table>

Source: Research ICT Africa (2008)

Network Generation

Cell-phone meta-data is used to generate four distinct empirical networks. In each of these networks, the vertices of the network are individual cell-phone subscribers and edges are added between subscribers who communicate with one another. However, as summarized in Table 4, these networks differ in (a) the amount of communication required for two vertices to be considered connected, and (b) the types of calls considered.

First, networks are differentiated by the threshold for connection. In the inclusive network specifications, nodes that exchange at least 3 SMSs or calls over the 7 month period of the data are treated as connected. This cutoff is chosen to exclude missed calls and single back-and-forth exchanges. In the more restrictive specifications, edges are only placed between individuals who have communicated at least once per month on average (7 times), thus capturing only strong
Table 4: Different Network Filters

<table>
<thead>
<tr>
<th></th>
<th>&gt;2 Calls or &gt;2 SMS</th>
<th>&gt;6 Calls or &gt;6 SMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-Days</td>
<td>Everyone</td>
<td>Strong Contacts</td>
</tr>
<tr>
<td>Weekend-Only</td>
<td>All Social</td>
<td>Strong Social</td>
</tr>
</tbody>
</table>

relationships. See Appendix C for details on the decisions to specify unweighted and undirected networks.

Second, networks are differentiated by the type of communications considered. The All-Days networks consider calls placed on any day, while Weekend-Only networks consider only calls placed on the weekend, restricting attention to calls likely to be non-commercial in nature.

For each of these networks, the 1% of users with the most connections are dropped from all networks to exclude large firms. This corresponds to dropping users with more than 170 and 91 distinct contacts for the All-Day networks, and 109 and 55 distinct users for the Weekend-Only networks.

Summary statistics for these different networks are presented in Table 5. It shows that changes in the communication threshold have a substantial effect on the number of users included, as many users are light users. When one moves from the most to the least inclusive criteria, the total number of users drops by almost half. Similarly, the restriction to weekend connections also causes a substantial reduction in the number of users, suggesting many phones are used primarily for business.

Theoretical guidance on which of these network filters is best does not exist. Theories of how network fragmentation may impede the ability of citizens to hold politicians accountable – like information diffusion models – are agnostic about the nature of connections. Similarly, models that suggest fragmented networks impede social sanctioning sometimes focus on sanctioning among friends and family (like Jackson, Rodriguez-Barraquer and Tan (2012)), but

Note that the restriction of attention to calls made in the mornings and evenings noted in Section 2 for geocoding users is not used in network generation. That restriction is only meant to ensure so inferences about one’s place of residence are not influenced by calls made at work.
sometimes suggest that sanctioning may also occur in commercial relationships (e.g. Fearon and Laitin (1996); Larson (2017b)). Rather than make a theoretically unfounded decision about which network to use based on which generates the best-looking results – the type of ex-post decision that has been implicated as a potential threat to the integrity of empirical social science research (Ioannidis, 2005; Open Science Collaboration, 2015) – this analysis presents results for all four of these networks in parallel.

4 Measure Validation

This section provides a number of descriptive statistics about CPM-generated network communities. It shows that these communities have numerous properties that comport closely with intuitive concepts of politically and socially salient networks, providing assurances that later analyses rest on a solid foundation.

Community Sizes

Table 6 presents the subscriber-weighted distribution of network community sizes. It shows that the median network community consists of 24-42 individuals for the four network specifications, though all networks show significant right-skew. This is most pronounced for the All-Days > 2 network where the largest communities are more than 100 times larger than in any other network. This may suggest that the network is insufficiently filtered, although as

---

14For the duration of Section 4, results are presented at the resolution selected by the atheoretic criterion for simplicity of presentation; results in formal analyses in later Sections are presented at all resolutions.
Table 6: Population-Weighted Community Statistics

<table>
<thead>
<tr>
<th></th>
<th>Mean Size</th>
<th>Median Size</th>
<th>Largest Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-Days, &gt;2</td>
<td>8,058</td>
<td>34</td>
<td>88,724</td>
</tr>
<tr>
<td>All-Days, &gt;6</td>
<td>50</td>
<td>42</td>
<td>507</td>
</tr>
<tr>
<td>Weekend-Only, &gt;2</td>
<td>45</td>
<td>38</td>
<td>401</td>
</tr>
<tr>
<td>Weekend-Only, &gt;6</td>
<td>26</td>
<td>24</td>
<td>195</td>
</tr>
</tbody>
</table>

previously noted, there is limited theoretical guidance to determine what network filter is most appropriate. Community size histograms can be found in Appendix H.

**Spatial Distributions of Communities**

The CPM assigns subscribers to network communities based only on calling patterns and does not take into account physical locations, so the spatial distribution of network communities is epiphenomenal and can be used as a test of measure validity.

Figure 1 plots the spatial distribution of 8 network communities selected at random from the 100 largest communities from the Weekend-Only > 2 network. The Figure shows a heat map where redness indicates the share of network community members living in a location.\(^{15}\) As the Figure shows, most communities consist either of (a) a group in a single city, or (b) a relatively concentrated group in a rural locality with a small diaspora located in a nearby urban center.

This pattern of rural communities linked to an urban diaspora – especially diaspora in Lusaka or the Copperbelt – is consistent with known patterns of rural-to-urban migration in Zambia. As of 2010, rural-to-urban migrants made up just over 5% of Zambia’s population, and of these 624,000 individuals, more than half (53%) lived in Lusaka and 30% lived in the Copperbelt [Zambian Central Statistics Office, 2013, p. 8-12]. Similar patterns can also be seen in other networks, as shown in Appendix I.

This pattern is also corroborated by spatial calling patterns. 61% of the average user’s

\(^{15}\)For simplicity, users are assumed to live at the centroid of the Second-order Thiessen Polygon. For more on geocoding, see Appendix F.
Figure 1: Spatial distribution of 8 random communities from the Weekend-Only $> 2$ Connections network. Figures show heatmaps of the geo-referenced locations of members of 8 network communities selected at random from the 100 largest communities from the weekend $> 2$ call network. The darkness of shading is proportional to the share of community members geo-located within a given area. The plots are relatively representative of the two main types of spatial distributions in the data: either (a) a large, densely clustered communities in a rural area and smaller diaspora communities located in one or more of Zambia’s urban centers (e.g. top right), or (b) hyper-urban communities located almost entirely within major cities (e.g. bottom left).
contacts are located within 20km of their location.\textsuperscript{16} This estimate likely underestimates the number of contacts in close proximity\textsuperscript{17} but even so, it does suggest that urban migrants are a significant part of communication networks.

The overall distribution of subscribers also suggests accurate geo-referencing. For both Constituencies and Wards (a sub-Constituency administrative unit, N=\textasciitilde1,420), the correlation between number of subscribers and population is quite high – 0.78 at the Constituency level and 0.81 for Wards – and population shares are approximately in line with PT market share (See Appendix J for scatter plots).

Additional evidence of measure validity can be found in the structure of network communities in Zambia’s Western Province. As detailed in Appendix K, Western Province has stood apart from the rest of Zambia since the British South African Company first turned it into a semi-autonomous protectorate. Today, it has self-governance rights, and is home to multiple secessionist political parties. In light of this history, one might expect network communities in Western Province to be unusually internally oriented.\textsuperscript{18} This can be tested using a measure of national integration for each of Zambia’s 9 Provinces.

National Integration is defined as the degree to which residents of a Province belong to network communities that include people from other Provinces. A Province where everyone is assigned to network communities with fellow residents would receive a score of 0; a Province where everyone belonged to a community consisting entirely of residents of other provinces would receive a score of 1. (The precise formula can be found in Appendix L).

Table\textsuperscript{7} presents these National Integration scores for each Province in Zambia for each network type, as well as an average across all four networks. As anticipated, Western Province has an extremely low national integration score. Indeed, only Zambia’s two urban centers of Lusaka and Copperbelt are more internally-oriented, despite Western Province being located relatively

\textsuperscript{16}Again assuming each user lives at the centroid of their Second-Order Thiessen Polygon for simplicity.

\textsuperscript{17}This is the case for two reasons. First, assuming all users are located at the centroid of their Thiessen Polygon amounts to assuming that population is uniformly distributed in space; in reality, users tend to cluster, and these clusters are often proximate to one another. And second, measurement error in geo-referencing will almost always lead to increases in the estimated distance between points.

\textsuperscript{18}The author is indebted to Pierre Englebert for this suggestion.
Table 7: National Integration Scores by Province

<table>
<thead>
<tr>
<th>Province</th>
<th>Avg Score</th>
<th>All-Days, &gt;2</th>
<th>All-Days, &gt;6</th>
<th>Weekend-Only, &gt;2</th>
<th>Weekend-Only, &gt;6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copperbelt</td>
<td>0.214</td>
<td>0.148</td>
<td>0.243</td>
<td>0.245</td>
<td>0.222</td>
</tr>
<tr>
<td>Lusaka</td>
<td>0.303</td>
<td>0.222</td>
<td>0.337</td>
<td>0.338</td>
<td>0.315</td>
</tr>
<tr>
<td><strong>Western</strong></td>
<td><strong>0.308</strong></td>
<td><strong>0.193</strong></td>
<td><strong>0.345</strong></td>
<td><strong>0.354</strong></td>
<td><strong>0.339</strong></td>
</tr>
<tr>
<td>Eastern</td>
<td>0.312</td>
<td>0.214</td>
<td>0.347</td>
<td>0.352</td>
<td>0.337</td>
</tr>
<tr>
<td>Southern</td>
<td>0.315</td>
<td>0.202</td>
<td>0.353</td>
<td>0.359</td>
<td>0.345</td>
</tr>
<tr>
<td>Northwestern</td>
<td>0.343</td>
<td>0.223</td>
<td>0.385</td>
<td>0.390</td>
<td>0.372</td>
</tr>
<tr>
<td>Northern</td>
<td>0.402</td>
<td>0.286</td>
<td>0.442</td>
<td>0.449</td>
<td>0.433</td>
</tr>
<tr>
<td>Central</td>
<td>0.415</td>
<td>0.290</td>
<td>0.461</td>
<td>0.465</td>
<td>0.443</td>
</tr>
<tr>
<td>Luapula</td>
<td>0.440</td>
<td>0.292</td>
<td>0.493</td>
<td>0.498</td>
<td>0.476</td>
</tr>
</tbody>
</table>

close to both these urban regions. As with the spatial distribution of network communities, this strongly suggests that network communities are capturing something closely related to what we intuitively think of as communities.

5 Network Fragmentation and ELF

Having established a measure of social network fragmentation in Section 2 and validated the measure in Section 4, this Section turns to the core substantive topic of this analysis: the relationship between ELF and network fragmentation.

To test the relationship between network fragmentation and ELF, network fragmentation \(NF_{i,p},(γ)\) for Constituency \(i\) in Province \(p\) calculated with resolution parameter \(γ \in (0, 1)\) is regressed on a set of constituency controls \((X_i)\), Province fixed-effects \((φ_p)\) and ELF \((ELF_i)\):

\[
NF_{i,p},(γ) = ELF φ + X_i δ + φ_p + ε_i
\]  

Constituency controls \(X_i\) consist of a set of controls to address sources of potential variation in the fidelity with which cell-phone meta-data captures the true structure of social networks,
including population density, share of the constituency that is rural, share of residents who are subscribers with the Partner Telecom, and dummies for the urban centers of Lusaka and the Copperbelt. In addition, province fixed effects are included to ensure that the variation examined exists among constituencies within each region and that results are not overly influenced by large regional differences in patterns of cell-phone usage or other unobservable factors, a distinct possibility in a country as regionally diverse as Zambia.

**Measuring ELF**

The measurement of ELF is necessarily sensitive to how ethnic groups are enumerated, and variations in enumeration choices can lead to substantial variations in measures (Posner, 2004). This is also complicated by variation in the salience of different ethnic identities over time in Zambia, as discussed in Appendix A. As data for this analysis comes from 2010, ELF is calculated with respect to the dimension of identity that has consistently been most political salient since the mid-1990s – what Posner (2005) terms *linguistic group* identity. As described in Appendix B, ELF is calculated according to a 5-fold *linguistic group* taxonomy using data from the Zambian census.

Note that *linguistic group* identity is not defined by one’s language. Most Zambians identify with one of ~ 72 tribes, each with its own language. However, certain tribes – generally those whose languages share a similar linguistic heritage – tend to also co-identify. To differentiate between these two (nested) levels of ethnic identity, Posner uses the terms *tribal* identity and *linguistic group* identity. However, it is important to emphasize that these labels are academic.

---

19 More precisely, the specification includes log population, log area, and log number of subscribers, which, given additive separability of logged ratios, is equivalent to the described parameters. Lusaka and Copperbelt dummies include core urban constituencies and immediate neighbors. *Lusaka dummy* includes constituencies of Katuba, Kafue, Chilanga, Chongwe, Chawama, Kabwata, Kanyama, Lusaka Central, Mandevu, Matero, and Munali. *Copperbelt dummy* includes constituencies of Chililabombwe, Chingola, Nchanga, Kalulushi, Chimwemwe, Kamfinsa, Kwacha, Nkana, Wusakile, Luanshya, Roan, Kankoyo, Kansanshi, Mufurila, Kafulafuta, Masaiti, Bwana Mkubwa, Chifubu, and Ndola.

20 There is some dispute about whether each tribe has its own language or whether some are dialects of one another. As pointed out by Kashoki (2017), however, the distinction is slippery at best.

21 These identities are nested, like identifying as both Irish and White, or Chinese and Asian.
constructions – both types of identity are referred to as “tribal” by Zambians (Posner, 2005, p.115). Moreover, as suggested by the fact that most tribes have their own language, the terms are obviously imprecise. Indeed, two people who belong to the same linguistic group need not speak the same language – they need only identify with tribes whose languages have similar linguistic roots.

In addition, individuals who identify with different linguistic groups often share a language due to high levels of bilingualism. Many Zambians speak one language in the home (often their tribal language) and another language for commercial transactions and social interactions (Laitin, 1992). Indeed, in the 1990 Zambian census, fully 25% of people who spoke Bemba, Nyanji, Tonga, or Lozi spoke it as a second language (Posner, 2005, p. 60).

Inter-intelligibility across linguistic groups is further facilitated by the existence of a *lingua franca* in each region. These *linguae francae* are officially recognized languages, are commonly used in media and commerce, are approved for use by civil servants, and, until recently, primary school instruction was required to be in either English or each region’s designated *lingua franca*.\(^{22, 23}\)

As such, while linguistic group is a convenient label for an ethnic taxonomy that is well understood among Zambians, it is measuring an identity that is much deeper than language and only somewhat related mutual intelligibility.

**ELF and Network Fragmentation Results**

Figure\(^2\) shows the partial correlation of ELF and network fragmentation for the full sample of Constituencies. Each estimate in the Figure is the $\beta$ coefficient from a regression of network

\(^{22}\)The one outlier in this pattern is Northwestern, which has three official *linguae francae* – Kaonde, Luvale, and Lunda. Other regions are restricted to instruction in a single *lingua franca*: Bemba in Copperbelt, Northern, Luapula, Kabwe (Urban), Mkushi and Serenje districts; Nyanji in Eastern, and Lusaka region; Tonga in Southern, Kabwe (Rural) and Mumbwa District; and Lozi in Western, Livingstone (Urban) (Kashoki, 2017b).

\(^{23}\)In addition, while distinct, these *linguae francae* are closely related, as all belong to the Bantu language family. Moreover, in a test of mutual intelligibility among school children, Kashoki (2017a) found that Bemba, Tonga, Lozi, and Nyanja speaking children understood about 30% of the content of passages read in other languages. These languages were also found to have an overlap of about 30-45% with one another in basic vocabulary.
fragmentation on ELF using the Equation 2 where network fragmentation is calculated using the value of $\gamma$ indicated on the x-axis. Moving from left to right along the x-axis corresponds to moving from a small number of large, loosely connected communities to lots of very small, dense communities. Coefficients have been normalized so that point estimates can be interpreted as the impact on network fragmentation (in standard deviations) of a movement of ELF from 0 (no fractionalization) to 1 (full fractionalization).

Three features of the Figure are worth noting. First, in nearly all networks and at all values of the resolution parameter, the point estimate for the partial correlation between ELF and network fragmentation is positive, as predicted by the network-proxy hypothesis. Second, these positive correlations are also statistically significant across a substantial range of resolution parameter values. Third, the estimated effect is also substantial – in the statistically significant estimates, a movement from non-fractionalized to fully-fractionalized constituencies appears to correspond to a relatively consistent 1-2 standard deviation increase in network fragmentation across all networks.

Finally, the ranges of $\gamma$ over which the relationship tends to be strongest correspond to network communities whose size comports well with our intuition of what scale of network structure might matter for political effectiveness. At $\gamma = -10$, average community sizes are approximately 2,000-30,000 people across network specifications, which roughly corresponds to the size of a large political rally or protest. The question of whether these are the resolutions at which the network-proxy would predict the correlation to be strongest, however, is addressed more systematically in Section 7.

It is also interesting to note that these community sizes are substantially smaller than the sizes of the ethnic groups in question, and are even smaller than the lower-level of ethnic identity.

---

24 In the limit, moving a little to the left of the plot corresponds to placing all nodes in one large community and moving a little past the right side corresponds to placing every node in its own community.

25 Results without province fixed effects (shown in Appendix N) show statistically significant estimates of a negative correlation for very small communities sizes ($\gamma > -5$, where median network communities consist of less than 24-42 individuals). As previously noted, however, province fixed effects are theoretically preferable given the regional heterogeneity of Zambia.

26 At $\gamma = -10$, population weighted average community sizes are 29,838 for All-Days, > 2 Calls, 17,008 for All-Days, > 6 Calls, 11,736 Weekend-Only, > 2 Calls, 2,091 for Weekend-Only, > 6 Calls.
groups in Zambia, tribes. This makes clear that while ethnically homogenous communities have less fragmented networks, this is not simply due to clean partitions along ethnic lines. Ethnic divisions clearly contribute to network fragmentation, but cannot alone explain network structure alone.

**Network Fragmentation and ELF**

Constituency-Level, By Resolution Parameter and Filter

Figure 2: Partial correlation coefficients between ELF and network fragmentation plotted at different community detection resolutions, ranging from large, relatively inclusive network communities on the left end of the x-axis to small, dense communities on the right end of the x-axis. Point estimates in red are statistically significant at the 90% level. Coefficients have been normalized so point estimates can be interpreted as the impact on network fragmentation (in standard deviations) of a movement of ELF from 0 (no fractionalization) to 1 (full fractionalization). Standard errors are clustered at the Province level.

One other aspect of these results worth noting is that the resolution parameter values at which ELF and network fragmentation are most correlated do not include the value of $\log(\gamma)$ that would be selected by the atheoretic criterion – approximately $-5$ in most graphs. Indeed, if one were to conduct this analysis using only the default value of $\gamma$ from an out-of-the-box
CPM package, one would erroneously conclude that ELF and network fragmentation are uncorrelated, illustrating the dangers of over-reliance on myopic statistical criteria when importing computational algorithms into the social sciences.

The positive association between ELF and network fragmentation is even more notable among rural constituencies. Figure 3 replicates Figure 2 for the 104 constituencies with at least 70% rural populations. The Figure shows that in the rural sample, both the magnitude of the ELF - network fragmentation relationship is larger and also the coefficients are statistically significant across a larger range of resolution parameters. 27

This result appears relatively robust. As shown in Appendix O, tightening the sample restriction to constituencies that are at least 85% rural strengthens the correlation significantly despite the decrease in sample size. Similarly, a persistent positive correlation remains when dropping all controls except the log number of subscribers and Province fixed effects. Results also do not appear to be driven by outliers, although they are not robust to exclusion of province fixed effects, which results in much noisier estimates and, as above, some estimates of negative correlations at very small network community sizes.

Two conclusions follow from these results. First, while additional work is needed to understand why the relationship only exists at certain resolutions – as addressed in Section 7 – the results in this section are consistent with the network-proxy hypothesis.

Second, these results suggest that the relationship between ethnicity and network fragmentation is weaker in cities than rural communities. This speaks to a broader literature on the effects of urban migration and modernization on ethnic identification, and on the question of whether urban migrants adopt class-based identities rather than ethnic identities. Much of this literature comes from the study of India, where evidence suggests that class-based cleavages have generally not been adopted (Weiner 1978; Thachil 2015). African cities have been less studied, but, to-date, results suggest a different dynamic. Melson (1971), for example, finds that

27 This sub-sample analysis, like the analysis across resolutions, was not a part of the author’s original research design. As such, the statistical significance of this post-hoc exploratory analysis should be interpreted with a measure of caution (Casey, Glennerster and Miguel 2012), though this limitation does not negate the value of the results (Laitin 2013; Olken 2015).
Figure 3: Partial correlation coefficients between ELF and network fragmentation for the 104 Constituencies that are at least 70% rural plotted at different community detection resolutions. Network community sizes range from large, relatively inclusive network communities on the left end of the x-axis to small, dense communities on the right end of the x-axis. Point estimates in red are statistically significant at the 90% level. Coefficients have been normalized so point estimates can be interpreted as the impact on network fragmentation (in standard deviations) of a movement of ELF from 0 (no fractionalization) to 1 (full fractionalization). Standard errors are clustered at the Province level.

Nigerian urban migrants seem to identify with both class and ethnicity. He finds a substantial share of workers support both a labor party and also an ethnic party at the same time. Similarly, in an audit survey of commercial interactions in an urban market in Lagos, Nigeria, Grossman and Honig (2015, p. 6) find “[n]on-coethnics who appear lower class are treated roughly the same as lower class coethnics,” while this is not the case for higher classes.

The findings from this analysis support the idea that urban migrants in African cities may be responding differently from those studied in India: while ethnicity is not unrelated to network
structure in cities, the relationship does appear to attenuate significantly in urban constituencies.

6 Network Fragmentation and Non-Salient Cleavages

The network-proxy hypothesis posits that ethnicity is political salient because the correspondence between ethnicity and network structure offers substantial organizational advantages to resultant political parties. A corollary of the hypothesis is that other cleavages – those which we do not see achieving political salience – must not offer this same advantage. To test for this, this section examines the relationship between network fragmentation and fragmentation along two other potential political cleavages that, despite many appealing attributes (like congruence with policy preferences), have not achieved salience in Zambia: employment sector and religious identification.

Figure 4 replicates the analysis from Section 5 (Specification 2) with economic fragmentation and religious fragmentation substituted for ELF. As the Figure shows, there is no systematic relationship between network fragmentation and fragmentation along these alternate potential political cleavages, consistent with the network-proxy hypothesis.

These results raise an obvious question: why is ethnicity uniquely correlated with network structure? Though answering this question rigorously is beyond the scope of this analysis, the network-proxy hypothesis does offer one potential explanation: British intervention and subsequent positive reinforcement.

As in many colonies, British administrative policies contributed significantly to the emergence of ethnicity as a politically salient identity. Indeed, as documented by Posner (2005, p.54), “[t]he reason Zambians identify themselves and others in tribal terms is because the institutions of the [British South African Company] and Colonial Office rule classified them in this manner and generated incentives for them to invest in these classifications.”

---

28 Religious fragmentation is an Herfindahl index of the share of Constituency residents who identify as Protestant, Catholic, Muslim, Hindu, Buddhist, Bahai, Other, or None in the 2010 Zambian Census. Economic fragmentation is a measure of fragmentation across the categories of Manufacturing, Farming/Fishing/Forestry, Services, and Mining in the 2010 Zambian Census. Summary statistics can be found in Appendix U.
Once ethnicity achieved salience, subsequent positive reinforcement may be responsible for the persistence of the alignment of ethnicity and network structure. The British created an incentive for politicians to mobilize voters along ethnic lines, but these mobilization efforts likely just reinforced ethnic networks. As a result, over time parties with ethnic constituencies developed ever-greater organizational advantages over parties that couldn’t take advantage of these ready-built networks. This would lead to more mobilization along ethnic lines, further reinforcing these networks in a cycle of positive reinforcement, eventually resulting in ethnicity’s uniquely strong relationship to network structure today.

Figure 4: Partial correlation coefficients between alternate potential cleavages and network structure. Network community sizes range from large, relatively inclusive network communities on the left end of the x-axis to small, dense communities on the right end of the x-axis. Point estimates in red are statistically significant at the 90% level. Coefficients have been normalized so point estimates can be interpreted as the impact on network fragmentation (in standard deviations) of a movement of employment sector fragmentation (left) or religious fragmentation (right) from 0 (no fractionalization) to 1 (full fractionalization). Standard errors are clustered at the Province level.

7 Public Goods and Voter Knowledge

The network-proxy hypothesis posits that ethnicity achieves political salience because it is correlated with network structure, and network structure matters politically because of its effect
on social dynamics like information diffusion or the ability of citizens to hold politicians accountable. If true, then (a) corollary outcomes like voter knowledge and quality of public goods should also be negatively correlated with network fragmentation, and (b) they should be negatively correlated with network fragmentation at the same resolutions that network fragmentation and ELF are most correlated.

This Section tests this by examining how the relationship between network fragmentation, voter knowledge, and the quality of public goods provision varies with the resolution parameter. It finds that (a) both voter knowledge and quality of public goods are generally negatively correlated with network fragmentation as predicted by theory, (b) these negative correlations are most statistically significant at resolutions where network fragmentation is positively correlated with ELF, as predicted by the network-proxy hypotheses, and (c) both voter knowledge and quality of public goods are most strongly correlated with network fragmentation at approximately the same resolution, suggesting the existence of a scale of maximal political significance.

**Public Goods and Knowledge Measurement**

Public goods are calculated using data on a number of public services from the 2000 and 2010 Zambian National Census 10%-sample micro-data, which include data on four essential public services: electrification, access to a protected water supply, child enrollment, and infant mortality. These measures are combined into a single measure by measuring change in each good from 2000 to 2010 for rural and urban households in each Constituency separately, calculating a weighted average of those values for each Constituency, normalizing those values, and extracting the first component of a Principal Component Analysis (PCA). A lengthier discussion of data used, tests of index validity, and motivation for use of an index can be found in Appendix R.

Political knowledge is measured using data from the 2009 Afrobarometer survey in which

---

29 Discussion of the motivation for selection of voter knowledge and public goods quality can be found in Appendix P.

30 The most statistically significant is also the substantively largest in nearly all cases.
1,200 Zambians were surveyed on a range of political issues in advance of the 2011 election. Political knowledge is operationalized as the share of political opinion questions to which respondents provided a response other than “Don’t Know/Haven’t Heard Enough.” For example, when asked “Now let’s speak about the present government of this country. How well or badly would you say the current government is handling the following matters, or haven’t you heard enough about them to say: Managing the economy?” a response is coded 1 if the respondent provides an evaluation or refuses to answer, and is coded as 0 if the respondent answers “Don’t Know / Haven’t Heard Enough”. This coding is applied to 15 different questions (a full list of included questions can be found in Appendix Q), and a “Knowledge Index” for each respondent is computed as the share of responses coded as 1. As the focus of this analysis is on network fragmentation measured at the level of electoral constituencies for the National Parliament, attention is restricted to questions pertaining directly to each Constituency’s MP or views of national government policy over which the MP has influence. The measure has an average value of 94.5% and a standard deviation of 0.08\(^{31}\)

**Public Goods and Knowledge Specifications**

The relationship between public goods and network fragmentation is estimated as:

\[
PublicGoods_{i,p,\gamma} = NF_{i,\beta} + X_i\delta + \phi_p + \epsilon_i
\]  

(3)

Where \(i\) is an index of Constituencies, \(\phi_p\) is a vector of Province fixed effects, and \(X_i\) is the set of Constituency-level controls used in Section \(^{5}\) Dummies are also included for whether the Constituency was represented by an MP with the ruling party (MMD) in either the first or second parliament of the 2000s (although results are robust to exclusion of the MMD controls and Lusaka and Copperbelt dummies).

\(^{31}\)The assumption underlying this measure is that exposure to information about government performance – either from direct experience or conversation – decreases the likelihood of a “Don’t know / Haven’t Heard Enough” response. The measure takes no position on the effects of information on attitude extremity or on the likelihood an individual will simply refuse to provide an answer (“Refuse to Answer” is a distinct code).
The relationship between voter knowledge and network fragmentation is estimated as:

\[ \text{PolKnowledgeIndex}_{j, t, p, y} = NF_{t, y} \beta + X_i \delta + Z_j \gamma + \phi_p + \epsilon_j \] (4)

where \( j \) is an index for individual survey respondents, \( \phi_p \) is a vector of Province fixed effects, \( Z_j \) are the individual-level controls including include gender, a PCA asset wealth index, urban/rural, and whether the respondent has completed primary school, and \( X_i \) are Constituency-level controls used in Section 5.

In both regressions, the coefficient of interest is \( \beta \) – the partial correlation of network fragmentation and the corollary outcome.

**Public Goods and Knowledge Results**

Figure 5 plots normalized point estimates of \( \beta \) and the 90th percentile confidence intervals for each value of the resolution parameter \( \gamma \in (0, 1) \). Recall that moving from left to right corresponds to moving from large, inclusive communities to smaller, more densely connected communities.

Several aspects of the Figure are notable. First, consistent with theory, network fragmentation is broadly negatively correlated with both public goods and voter knowledge – that is, the point estimates are generally negative. For public goods, all statistically significant estimates are negative in three of the four networks; for voter knowledge, all statistically significant coefficients are negative.

Second, the resolutions at which political knowledge, voter knowledge, and network fragmentation are most closely correlated are quite distinct from those at which network communities are most unlikely in a random graph. These results show that we may be looking at the wrong resolution of networks when we rely on the atheoretic criterion, and point to the importance of thinking about the relevant level at which to measure fragmentation before making claims about how social dynamics affect political outcomes. (See Appendix T for further evalu-
ation of differences between results at these most-correlated resolutions and resolutions chosen using an atheoretic criterion.)

Third, the resolution parameters that are most correlated are also very similar for the two social phenomena examined here (which are qualitatively different and come from different sources), as illustrated in Table 8. This strongly suggests that network fragmentation at these resolutions is robustly salient for social and political outcomes.

Table 8: Most Significant (Logged) Resolution Parameters

<table>
<thead>
<tr>
<th></th>
<th>Pol Knowledge</th>
<th>Public Goods</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-Days &gt;2</td>
<td>-7.50</td>
<td>-7.50</td>
<td>-7.50</td>
</tr>
<tr>
<td>All-Days &gt;6</td>
<td>-6.99</td>
<td>-7.99</td>
<td>-7.49</td>
</tr>
<tr>
<td>Weekend-Only &gt;2</td>
<td>-7.50</td>
<td>-9.50</td>
<td>-8.50</td>
</tr>
<tr>
<td>Weekend-Only &gt;6</td>
<td>-14.51</td>
<td>-12.02</td>
<td>-13.27</td>
</tr>
</tbody>
</table>

Correlation: .91

And finally, the resolutions at which public goods and voter knowledge are most correlated with network fragmentation are also close to the resolutions where ELF and network fragmentation have a positive and statistically significant correlation. Figure 6 replicates the plot of ELF and network fragmentation partial correlations from Section 5 with the addition of a solid vertical line at the average of the resolutions most correlated with voter knowledge and public goods. The Figure shows that regions where ELF and network fragmentation are especially well-correlated are relatively close to the resolutions most correlated with public goods and voter knowledge. This suggests not only that ELF and network fragmentation are consistently correlated at many resolutions – as illustrated in Section 5 – but also that these correlations occur at politically relevant resolutions, providing further support for the network-proxy hypothesis.

8 Conclusion

This work makes several important contributions to our understanding of the role of ethnicity and social networks in the politics of developing democracies, and also to the study of civil
society more broadly. Substantively, it provides clear evidence in support of the assumption underlying the network-proxy hypothesis: that ethnically fragmented communities have more fragmented social networks, and that this is not the case for alternate political cleavages that, tellingly, have not achieved political salience. In addition, it shows that voter knowledge and public goods tend to be negatively correlated with network fragmentation, providing further support for the network-proxy hypothesis.

The dynamics of ethnic politics in Zambian are very similar to those in other African countries, limiting concerns about external validity. The largest potential scope condition is that the dimension of ethnicity used in this analysis is related to language, as discussed in Section 5. As such, this correlation between network fragmentation and ELF may be more tenuous in communities where ethnicity is more divorced from language. As ethnicity and language are often intrinsically related, however, especially in Africa, even if this is the case, these findings are still likely to be informative for a large number of contexts.

These results also shed light on the results of other studies. For example, they are consistent with the Dionne (2015) finding that trust in behavioral games is higher among strongly connected individuals regardless of whether they are co-ethnics (suggesting it is networks and not ethnicity that drives trust dynamics). The finding from Dionne (2015) that trust is higher among non-co-ethnics when partners are not close friends is harder to rationalize with the network-proxy hypothesis, and may be due to either something idiosyncratic about the social network within the village studied or as noted by Dionne (2015) it may suggest that intra-village dynamics diverge from social dynamics at larger scales. This analysis is also consistent with the empirical finding from Larson and Lewis (2017) that ethnically fragmented villages diffuse information less efficiently, although less consistent with the finding that ethnically fragmented communities have higher link density. Further research is needed to understand whether this

---

32 The very fact that our standard measure of ethnic fractionalization is a measure of ethno-linguistic fractionalization underlines how intertwined these two dimensions are in our study of ethnic politics.

33 Suggesting this may be the case, Dionne (2015) found no evidence that reported close friends were any more likely to be co-ethnics than non-co-ethnics, an unusual finding given pervasive findings of homophily across a broad range of contexts (McPherson, Smith-Lovin and Cook, 2001).
divergence is due to the small sample size (two villages) in Larson and Lewis (2017), the scale of network structure being measured in Larson and Lewis (2017) (they look only at very local network properties and not at meso-scale properties like community fragmentation as analyzed in this paper), or whether the social process that the analysis measures is distinct from the processes inherent to political mobilization.

Methodologically, this analysis also offers several important advancements in how to address the problem of scale in a social science setting. As an analogous problem exists in many inductive algorithms, including all clustering algorithms (like text-as-data Topic Models), these innovations have implications beyond network analysis.

Finally, this analysis points to a number of potential avenues for future research. For example, the finding that the network-ethnicity relationship is weaker in urban communities than rural communities begs for further investigation, as an improved understanding of the mechanisms behind this attenuation may be of value to policymakers seeking to decrease the political salience of ethnicity. And more broadly, this analysis illustrates how new sources of big data on the day-to-day interactions of citizens can be leveraged to rigorously answer questions about social organization, civil society, and political outcomes. Since at least the works of Putnam, Leonardi and Nanetti (1993) and Almond and Verba (1989), scholars have posited that social capital and an effective civil society are crucial to good democratic governance because of their role in holding politicians to account. Yet most research in this area has been unable to systematically measure social capital. Numbers of NGOs or membership in formal organizations are often-used measures that fail to capture the richness of the theories that they aim to test. This analysis illustrates how newly available data like cell-phone meta-data can be leveraged to develop empirical measures that are as rich as the theoretical literatures we aim to test.
References


40


URL: [http://stanford.edu/ ejdemyr/docs/EKR-20151103.pdf](http://stanford.edu/ ejdemyr/docs/EKR-20151103.pdf)


URL: http://danhonig.info/sites/default/files/these_are_my_people_-_grossman_and_honig_2015.pdf


Figure 5: Partial correlation coefficients between the public goods index (top) / voter knowledge (bottom) and network fragmentation plotted at different community detection resolutions, ranging from large, relatively inclusive network communities on the left end of the x-axis to small, dense communities on the right end of the x-axis. Intervals are 90th percentile confidence intervals. The vertical dashed lines indicate the resolution selected by the atheoretic criterion. The vertical solid lines denote the resolutions at which the partial correlation coefficients are most significant for each specification. For readability, coefficients are normalized by their standard errors prior to plotting; as such, values correspond the t-statistic associated with each estimate. Plots with substantive magnitudes can be found in Appendix S. Standard errors are clustered at the Province level for public goods and survey-site level for voter knowledge.
Figure 6: This Figure plots partial correlation coefficients between ELF and network fragmentation at different community detection resolutions, ranging from large, relatively inclusive network communities on the left end of the x-axis to small, dense communities on the right end of the x-axis. 90th percentile confidence intervals. The vertical solid lines denote the resolution at which the partial correlation coefficients between network fragmentation and public goods / voter knowledge are most significant.
Appendix to
“Social Networks and the Political Salience of Ethnicity”

Nicholas Eubank

Date: August 17, 2018

Appendix Contents

Ethnicity in Zambian Politics ................................................................. page i
Calculating ELF .................................................................................. page iv
Network Specification ........................................................................ page v
Constant Potts Model ........................................................................ page viii
Optimization of the Constant Potts Model ........................................ page x
Geo-Referencing Cell-Phone Users ................................................... page xi
Public Goods Administration ............................................................ page xiv
Community Size Histogram .............................................................. page xv
Spatial Distribution of Communities ................................................ page xvi
Population and Geo-referenced Subscribers .................................... page xix
Western Region’s Secessionist Tendencies ....................................... page xx
Calculation of National Integration ................................................... page xxi
Computation of Weighted Fragmentation Measures ....................... page xxii
Robustness to Dropping Province Fixed-Effects .............................. page xxiii
Robustness of Rural-Sample Results .............................................. page xxv
Selection of Public Goods and Voter Knowledge ........................... page xxviii
Political Knowledge Questions ....................................................... page xxviii
Computing a Public Goods Index ..................................................... page xxx
Public Goods, Voter Knowledge, and Network Fragmentation .......... page xxxiii
Results By Resolution Selection Criterion ..................................... page xxxvii
Alternate Cleavages ........................................................................ page xli
Additional Acknowledgements ........................................................ page xli

A Ethnicity in Zambian Politics

Zambia’s political history can be divided into three periods (generally referred to as the First, Second, and Third Republics), and the nature and relative salience of ethnic identification has varied across these periods. From Independence in 1964 to 1972, Zambia was a multi-party democracy; from 1972 to 1991, Zambia was subject to the single-party rule of Kenneth Kaunda
and the United National Independence Party (UNIP) from 1972 to 1991; and finally from 1991 to the present day, Zambia has returned to being a multi-party democracy.

During the post-independence period, the rise of ethnicity as a salient political issue took President Kaunda – who expected Zambians to unite around a national identity – by surprise (Phiri, 2006). Despite winning Zambia’s first election by a landslide under the slogan “One Zambia, One Nation,” President Kaunda quickly found his administration embroiled in debates over ethnic representation. “Political struggles during the First Republic often took a regional or ethnic dimension and this continued to be a lasting feature of Zambian politics,” and the President was forced to address these issues through a deliberate policy of ethnic balancing (Rabe, 2016, p.179). Even this was challenging, however, and fights over the ethnic composition of the administration nearly destroyed the UNIP party, leading to the temporary resignation of President Kaunda in 1968 (Phiri, 2006; Sardanis, 2014). Eventually one of the chief protagonists in these struggles – Simon Kapwepwe – left the UNIP to form the United Progressive Party (UPP) opposition party in 1971 (Sardanis, 2014, p.71), stating that “The people of the northern part of Zambia – the Bemba-speaking people – have suffered physically... They have suffered demotions and suspicions because of my being Vice-President. I cannot sacrifice any longer these people.” (Phiri, 2006, p.149).

Faced with this growing discontent, President Kaunda amended the Zambian constitution to outlaw all opposition parties in 1972, moving Zambian into a period of one-party rule during which the nature of ethnic political competition in Zambia changed markedly. The vast majority of Zambians belong to one of roughly 70 tribes. Prior to colonialization, each of these tribes also corresponded to a distinct linguistic group. For a number of reasons pertaining primarily to colonial institutions, however, by the time of independence, members of most tribes identified with only a small number (four to seven, depending on the classification scheme) of consolidated linguistic groups. During the First Republic, when ethnic representation was contested, it was in terms of the representation of these ethno-linguistic categories.

During the Second Republic, however, this changed. The reason – as argued by Posner (2005) – is that these linguistic groups tend to be relatively concentrated in specific regions. Under one party rule, however, elections were never about determining who would hold national power, but rather which local group would get to elect the representative to the UNIP-led one-party state. In other words, “the institutions of the one-party state shifted the locus of electoral competition from the national to the local level, and this led to an increase in the salience of more localized ethnic identities.” (Posner, 2005, p. 195). Thus the specific ethnic identity that

---

\[\text{Drivers include a desire of missionary schools to minimize the number of languages of instruction, and subsequent education policies that formalized the dominance of certain languages. See Posner (2005) (pages 56 - 88) for a detailed discussion the factors driving this consolidations.}\]
was politically salient shifted, but as [Posner (2005)] makes clear ethnicity still remained at the forefront of political debate and voting.

Indeed, perhaps the only election in which ethnicity was not particularly salient was the first election in Zambia’s Third Republic. Under the banner of the Movement for Multi-Party Democracy (MMD), a broad coalition banded together to oust the UNIP party that had long held control during Zambia’s one party regime ([Posner 2005] p. 187). But this unity proved to be ephemeral. Factions within the party quickly organized along linguistic-group ethnic divisions ([Posner 2005] p. 188). Between 1991 and 1994, 13 ministers and deputy ministers and the Vice-President had resigned or been fired, leading to a concentration of Bemba politicians in leadership positions of the MMD and a rise in “fear of dominance of Bemba speakers in politics.” As a result, by 1994, the MMD “was no longer a national party enjoying the popular and legitimate support of the whole population as was the case in the pre-election time.” ([Osei-Hwedie 1998] p. 236)

Today, while parties occasionally align with specific ethnic groups (like the United Party for National Development and the Tonga ([Erdmann 2007])), in general most parties are multi-ethnic, but carefully manage (and debate) the distribution of influence of ethnic groups within parties (e.g. the United National Independence Party, Patriotic Front, and Movement for Multi-Party Democracy). Indeed, policies of “ethnic balancing” have remained a core electoral strategy – and source of internal strife – for nearly every ruling Zambian regime.
B Calculating ELF

As noted in body of this analysis, the measurement of ELF is necessarily sensitive to how ethnic groups are enumerated, and variations in enumeration choices can lead to substantial variations in measures (Posner, 2004). This analysis focuses on the most political salience dimension of ethnic identity in Zambia – linguistic groups.

Beyond selection of the relevant dimension of identity, however, calculation of ELF is also dependent on the number of linguistic groups enumerated. The Zambian Central Statistics Office (CSO), for example, identifies seven major language groups: Barotse, Bemba, Mambwe, North-Western, Nyanja, Tonga, and Tumbuka (Zambia 2010 Census of Population and Housing, 2012, p. 65) which collectively account for over 98% of Zambians.

Political scientists, by contrast, tend to focus on only five major groups: Bemba, Nyanja, Lozi/Baroste, Tonga, and “North-Western” (Posner, 2005; Osei-Hwedie, 1998). In the case of Posner (2005), this move to a 5-way taxonomy is accomplished by grouping the Tumbuka (3.3% of Zambia’s population) with the Nyanja, and grouping the Mambwe (1.3%) with the Bemba, arguing these groups represent salient political coalitions.

In light of the relative agreement between scholars of Zambian politics on the salience of these five ethno-linguistic coalitions (Posner, 2005; Osei-Hwedie, 1998; Gibson and Hoffman, 2013), this analysis relies on the five-fold taxonomy described above. However, note that the difference between the seven-fold and five-fold taxonomies is quite small given the size of the last two groups – at the level of electoral constituencies, the correlation in ELF computed using these different categorizations is 0.96. The residual 1% of individuals with alternate stated ethnicities are grouped under the label of “Other” and included in calculations of ELF as a sixth group, while those who do not state an ethnicity are excluded.

---

2“North-Western” is not a single language group per se, but rather a collection of language groups that co-identify. The North-Western region was never subject to the same colonial pressures to consolidate around a single language as other regions, and Posner argues this distinct history has driven the North-Western region to act analogously to other linguistic groups. “Zambians often refer to ‘North-Westerners’ as the fifth major ethnic group alongside the Bemba-speakers, Nyanja-speakers, Tonga-speakers, and Lozi-speakers. People from North-western Province also commonly identify themselves in such terms.” (Posner, 2005, p.119)

3Individuals are assigned to a linguistic group based on their response to their stated tribal ethnicity, with tribes grouped into aggregate linguistic groups based on CSO groupings from the documentation for the 1980 census. The only modifications to the CSO groups are for Tumbuka (who are grouped with the Nyanja-speaking group) and the Mambwe (who are grouped with the Bemba-speaking group) following Posner (2005).
C Network Specification

Motivation for Use of Unweighted Specification

The decision to generate an unweighted network and not take into account the frequency or duration of calls between individuals as weights – as has been done in some other studies like Onnela et al. (2007); Miritello et al. (2013) – is motivated by two considerations.

First, it is not clear that frequency of communication is necessarily a good indicator of the social importance of influence of a relationship. There is some evidence that within a type of relationship (e.g. among co-workers) frequency of communication in one electronic medium may be a good proxy for intensity of communication across all mediums (Haythornthwaite 2005), but it is far less clear whether this holds across types of relationships. For example, people in some industries may make more frequent calls to co-workers and business partners than to family members, but this does not necessarily mean that those relationships are more socially salient or influential. Indeed, in a survey of 40 US individuals who agreed to share phone records and fill out questionnaires about their connections, Wiese et al. (2015) finds that while call frequency and duration are a reasonable predictors of self-reported tie strength, “many people in all tie strength levels had very little communication.” (Wiese et al., 2015, p.5). In subsequent interviews about mis-classifications, Wiese concludes errors arise from several factors, including substitution into “in-person communication,” the fact “[f]amily is close regardless of communication,” and that “[o]ther participants used instant messenger, email, Skype, or SMS replacements such as WhatsApp to stay in touch with close contacts.” (Wiese et al., 2015, p.7).

Second, and even more importantly, even if frequency of calls is a reasonable proxy for connection strength in a developed-country context, frequency of calls is an especially problematic metric in a developing country context. This is because unlike in developed countries, the cost of phone calls – while very low – is non-trivial to the average Zambian. As a result, Zambians are especially likely to substitute away from phone calls and into face-to-face communications among geographically proximate contacts. Indeed, fear of omitting social ties among individuals who live in the same village is precisely the motivation for keeping the threshold for connection low during network generation – again, even in the more stringent networks, individuals are considered connected if they have exchanged just one texts or call per month.

---

4Indeed, (Haythornthwaite 2005, p. 125) concludes only “that media use within groups [emphasis mine] conformed to a unidimensional scale.”

5Though set in a US context, Motahari et al. (2012) shows that calling patterns among family members are qualitatively very different from calling patterns with other parties (among the Californians studied, calls to family members are more very frequent but much shorter than calls to other parties). The study does not report differences in total call times, but evidence of different usage patterns is consistent with the idea that the mapping from call frequency or duration to tie-significance may vary across types of connections.
In an ideal world, of course, it would be possible to at least partially correct for this by creating a measure of tie strength based on “frequency of calls controlling for distance” measure. Unfortunately, however, the structure of the data used in this analysis do not allow for such a measure – while the data includes call information and unique identifiers for all users, information on which antenna tower routes each call is only available for PT subscribers. This precludes geo-referencing non-PT subscribers or creating a “distance-adjusted” measure of connectivity.

However, even a distance-corrected measure would not compensate for a wealth effect. Due to the non-trivial cost of communications, using duration or frequency of calls as a metric for social connection would also privilege connections among affluent users, as they are likely to place calls more frequently and communicate for longer periods.

In an effort to strike a compromise between minimizing measurement error and learning about how different types of network ties may differ in their importance to network processes, this analysis does subset the analysis along two dimensions. First, it presents results when one limits results to weekend and evening calls – likely subsetting to friends and family to the exclusion of co-workers. And second, it examines two different inclusion thresholds – a minimum number of calls two parties must exchange to be considered connected.

This thresholding offers a number of advantages over the use of, say, length of calls as a measure of the social significance of a connection. First, it leverages the extensive margin but not the intensive margin. These helps obviate problems caused by the likelihood cost-conscious users avoid long conversations with geographically-proximate peers. So long as some communications take place – say, coordinating a meeting – connections among geographically proximate parties will still appear in the network. And second, while measurement error will still impact inferences of connections, this measurement error will be limited to ties whose intensity is in the immediate proximity of the threshold.

**Motivation for Use of Undirected Specification**

The decision to treat connections as undirected also has a two-fold motivation. First, information exchange in phone communications is inherently bi-directional, regardless of who places the call. Second, and perhaps more importantly, the direction of a call can be surprisingly difficult to establish in the Zambian context. In Zambia, the cost of a call is borne by the person placing the call. As a result, many users engage in the practice of giving more affluent contacts a missed call (they call, let the phone ring once, then hang up) as a signal that they would like the more affluent contact to call them back, allowing the more affluent party to be billed for the call. These missed calls do not appear in the data however (call detail records are primarily collected for billing purposes, and missed calls aren’t billed, meaning a bi-directional...
relationship may appear (in the data) to be uni-directional.
D Constant Potts Model

Let $N(V, E)$ be a social network where $V$ is the set of vertices (nodes) in the network and $E$ is the set of edges (links) between vertices. In network-theoretic terms, the objective of a community detection algorithm is to find a partition $\sigma$ of the social network $N$ where every individual vertex $i \in V$ of the network is assigned to exactly one group (where vertex $i$’s assignment is denoted $\sigma_i$).

Finding a partition requires two things: a formal measure of the quality of a partition, and an algorithm which optimizes that measure. To comport with intuition and theory, a measure of partition quality should be (a) increasing in the number of connections among individuals within a community, and (b) decreasing in the number of subscribers who are in the same community but are not connected. More specifically, for a given partition of the network $\sigma$ and summing across all subscribers $i, j \in V$, this intuition can be captured with the following objective function:

$$
\mathcal{H}(\sigma) = -\sum_{i,j \in V} [aA_{ij} - b(1 - A_{ij})] \delta(\sigma_i, \sigma_j)
$$

(1)

where $A$ is an adjacency matrix and $A_{ij}$ is equal 1 if $i$ and $j$ are connected and 0 otherwise; $\delta(\sigma_i, \sigma_j)$ is an indicator function that takes on a value of 1 if subscriber $i$’s community $\sigma_i$ is equal to subscriber $j$’s community $\sigma_j$ and zero otherwise; and $a$ is the positive weight of a connection between subscribers in the same community, $b$ is a penalty weight for two subscribers in the same community not being connected. Note that for a network with a fixed number of connections (edges), a partition in which communities have many internal connections will necessarily have few connections between communities.

This objective function can be further simplified by noting that the absolute magnitudes of $a$ and $b$ do not affect optimization, only their relative magnitudes. Setting $a = 1 - \gamma$ and $b = \gamma$ (where $\gamma \in (0, 1)$), this becomes:

$$
\mathcal{H}(\sigma, \gamma) = -\sum_{i,j \in V} [A_{ij} - \gamma] \delta(\sigma_i, \sigma_j)
$$

(2)

also known as the Constant Potts Model (CPM) (Traag, Van Dooren and Nesterov, 2011).

Note that as $\gamma$ dictates the relative cost and benefit of adding a vertex to an existing community when that vertex is connected to some members of the community but not others, the value of $\gamma$ determines the density of communities that maximize the objective function. In particular, adding a vertex to an existing community will increase $\mathcal{H}$ by $(1 - \gamma)$ for each connection the new vertex shares with a member of the existing community, but decrease $\mathcal{H}$ by $\gamma$ for each

viii
member of the existing community with whom it is not connected. As a result, the higher the value of $\gamma$, the less likely a marginal vertex is to be added to a community, and the smaller and more densely connected the resultant network communities will be. For this reason, $\gamma$ is often referred to as the \textit{resolution parameter} of a CPM function.

Because $\gamma$ determines the size and density of network communities, the choice of $\gamma$ is directly tied to the problem of scale described in the introduction. The choice of a large value of $\gamma$ (high resolution) will generate very small, very densely connected communities (like a group of households within a village), while a low value of $\gamma$ will give rise to large but more loosely connected communities (more like a group of villages within a region).\footnote{This flexibility differentiates the CPM objective function from other approaches – like modularity optimization – which are subject to fundamental resolution limits that preclude them from detecting communities below a given size \cite{Fortunato2007, Traag2011}.}
Optimization of the Constant Potts Model (CPM) community detection algorithm is accomplished using the Louvain algorithm. The Louvain algorithm is a greedy agglomeration algorithm for community detection, and consists of two stages iterated indefinitely. The algorithm begins with each vertex assigned to its own community. In the first stage, the algorithm iterates over vertices in the network in a random order and, for each vertex $i$, measures the improvement in the CPM objective function that would come from putting $i$ in the community of one of $i$’s neighbors. If merging $i$ with a neighbor can improve the CPM score for the network, $i$ is merged with the neighbor that most maximizes the CPM score. If no merges improve the CPM score, $i$ is left in its own community. The first stage iterates over vertices (potentially revisiting nodes) until it reaches a local maximum where no CPM-score-improving merges remain. In the second stage, a new weighted graph is created by merging all the vertices in each community into a single vertex where the new vertex has edge weights equal to the sum of the edge weights (all edges are assumed to have weight 1 if the input graph is unweighted) of the members of the community from which it is created. The first and second stages are then repeated until convergence.

While the Louvain algorithm is a greedy algorithm (Blondel et al., 2008) and is not guaranteed to find a global maximum, in practice is has been shown to perform extremely well (Lancichinetti and Fortunato, 2009), and order of vertex consideration does not appear to substantially impact results (Blondel et al., 2008). Moreover, it is computationally tractable even on very large networks, which cannot be said for most alternative methods, such as simulated annealing. Indeed, it is only because of the scalability the Constant Potts Model (optimized via the Louvain algorithm) that this project is possible, as other measures of network structure sometimes used in similar applications – like point connectivity (Moody and White, 2003) or random-walk average commute times (Yen et al., 2005) – are computationally infeasible on networks of this size. Laplacian-matrix based methods for computing average commute times do not scale well, and even simulation-based methods are problematic as the average number of edges a random walker must traverse to get from one randomly selected individual to another on a network with millions of vertices and tens of millions of edges make simulation-based estimation computationally intractable.
Geo-Referencing Cell-Phone Users

Geo-referencing is accomplished by taking advantage of the fact that most calls are routed through the cell tower closest to the user. To take advantage of this fact, the first step in geo-referencing users is to estimate the geographic areas served by each antenna tower. This is accomplished through the use of Thiessen polygons. A Thiessen polygon for an antenna tower \(a\) is the set of all points that are closer to \(a\) than any other antenna tower. Or, more formally, for an antenna tower \(a \in A\) and a point in space \(p \in P\), the Thiessen polygon associated with antenna tower \(a\) is given by:

\[
T(a) \equiv \{ p : \| a - p \| < \| a' - p \| \quad \forall a' \in A \text{ where } a' \neq a \}
\]

To this basic definition, however, it is also necessary to bound the range of an antenna tower. This paper assumes that a tower’s coverage is bounded by a distance of 35km, the technical limit of standard GSM antennas. Thus the more accurate definition is:

\[
T(a) \equiv \{ p : \| a - p \| < 35km \text{ and } \\
\| a - p \| < \| a' - p \| \quad \forall a' \in A \text{ where } a' \neq a \}
\]

An illustration of this type of Thiessen polygon is given in Figure (a) and (b), which provides a graphical illustration of the geo-referencing process.

Since the Thiessen polygon defines the geographic area for which a given tower is closest, and calls are most likely to be routed through the nearest tower, it follows that if most of cell-phone subscriber \(i\)’s calls are routed through antenna tower \(a\), subscriber \(i\) likely lives in the Thiessen polygon associated with antenna tower \(a\): \(T(a)\). However, only taking into account a user’s most used tower leaves out substantial information. By assuming that a user’s most used tower is the closest tower to the user, and further assuming that the second most-used tower is the second most used, estimates of where a subscriber lives can be further refined.

This is accomplished using second-order Thiessen polygons. A second-order Thiessen polygon is the set of points for which a given antenna tower \(a_1\) is closer than any other tower, and a second tower \(a_2\) is closer than any other tower except for \(a_1\). More formally, a second-order
Thiessen polygon associated with a primary tower $a_1$ and a secondary tower $a_2$ is defined as:

$$T(a_1, a_2) \equiv \{ p : \|a_1 - p\| < 35km \quad \& \quad \|a_1 - p\| < \|d' - p\| \quad \forall d' \in A \quad \text{where} \quad d' \neq a_1 \quad \& \quad \|a_2 - p\| < \|a'' - p\| \quad \forall a'' \in A \quad \text{where} \quad a'' \neq a_1, a_2 \}$$

Note that as illustrated in Figure [1](c), second-order Thiessen polygons constitute subsets of first order Thiessens. Further, to preclude the possibility that a user’s second-most used tower actually corresponds to another home or location the subscriber frequents, attention is restricted to the second-most used tower from the set of towers close enough to the most-used tower to ensure the set of second-order Thiessens is non-empty.\(^8\)

Once computed, these Second-Order Thiessen Polygons are used as estimates of where cell-phone users reside. A user whose most used cell-phone tower is Tower 10 and second most used tower is Tower 11, for example, is assumed to live in the Thiessen $T(10, 11)$ – the Second-Order Thiessen Polygon defined as the set of all points for which Tower 10 is closest and Tower 11 is second-closest.

Inferences about each users most-used tower are based on the total number of calls placed by a user before 8am and after 6pm, when the user is most likely to be in their residence (rather than their place of work). No other filters (such as the filters used in social network generation, like considering only communications along ties with a minimum number of calls or SMS messages) are applied.

\(^8\)In theory, further iterations of this algorithm are possible (third or fourth degree, for example), but in practice this is often not practical. In many rural areas, antenna towers are arranged along major roadways in a single-dimensional line. As such, considering a third or fourth tower simple means making minor adjustments to the placement of individuals along the roadway without any improvement in distance from roadway.\(^9\) Moreover, the number of Thiessens increases exponentially as higher orders are considered, introducing substantial computational difficulties.
Computing Second-Order Thiessen Polygons

(a) (b)

(c) (d)

Figure 1: Geo-referencing proceeds in four steps. In (a), cell-phone antenna towers are geo-referenced. In (b), Zambia is partitioned into regions based on the identity of the closest cell-phone antenna tower. At all points within $T(3)$, for example, Antenna 3 is the closest cell-phone tower. These regions also are limited to a maximum distance of 35km from an antenna tower. These are First-Order Thiessen Polygons. In (c), these regions are then subdivided based on the identity of the second closest tower. For all points within $T(3, 4)$, for example, Antenna 3 is the closest antenna tower, and Antenna 4 is the second closest antenna tower. These are Second-Order Thiessen Polygons. Finally, in (d) these Second-Order Thiessen Polygons are intersected with census boundaries, as discussed in Section 2.
G Public Goods Administration

On paper, the National Assembly and a sub-national political body called a Local Council share *de jure* responsibility for delivering primary care, health protection, and roads, and Local Councils are solely responsible for provision of utilities like water and electricity, while the government has *de jure* authority over education, police, and sea and air ports. In reality, however, while some services are provided entirely by the National Government, almost no public goods are under the unique purview of Local Councils. Even in water and electricity, national authorities often play a large role. In water policy, for example, the National Authority for Water and Sanitation Council (NWASCO), Ministry of Energy and Water Development (MEWD), District Water, Sanitation and Health Education Committees (D-WASHE), and Rural Water Supply and Sanitation Unit (RWSSU) have regulatory authority over the sector and are deeply involved with service provision (in part due to the limited capacity of some Local Councils) (USAID 2008). The Rural Electrification Authority (REA) and the Energy Regulation Board of Zambia (ERB) (Haanyika 2008) play a similar role in electricity provision. Indeed, the National government also has a specific instrument designed to allow it to achieve administrative but not political decentralization when it desires – District Commissioners, who are appointed by and work for the national government, but have identical geographic jurisdictions to Local Councils. This has led some cynical observers to suggest that the “decentralization aspect of [this policy] is just wishful thinking” (Sardamis 2014 p.240).
Figure 2 plots histograms of the network community sizes summarized in Section 4.

Figure 2: Distribution of network community sizes. Each plot shows a histogram of network community sizes under different network specification.
I Spatial Distribution of Communities

Figure 3: Spatial Distribution of 8 Random Communities, Most Inclusive Network
Figure 4: Spatial Distribution of 8 Random Communities, All > 6

Community Size: 348 users

Community Size: 386 users

Community Size: 289 users

Community Size: 301 users

Community Size: 283 users

Community Size: 245 users

Community Size: 231 users

Community Size: 273 users
Figure 5: Spatial Distribution of 8 Random Communities, Weekends > 6
Figure 6: Scatter plots of census population and number of georeferenced subscribers. Note that random noise has been added to census populations to protect commercially sensitive information. Actual correlations without noise presented below each figure.

J Population and Geo-referenced subscribers

Figure 6 plots the relationship between population and number of subscribers for Constituencies and Wards. Note that noise has been added to these plots to protect against de-anonymization, reducing the apparent fit.

Careful readers may be concerned that as population density is used as an input to subscriber geo-referencing, a strong correlation between subscribers and population is unsurprising. However, note that population density is only used to determine the proportion of subscribers allocated to each district when a Thiessen cuts across district boundaries; at no time is population density used to determine the absolute number of subscribers in a region, and population density only affects the allocation of subscribers when a Thiessen cuts across district boundaries.

\[ \text{(xvii)} \]
\[ \text{The number of subscribers per district is commercially sensitive and without noise, census populations could be used to identify Constituencies.} \]
Western Region’s Secessionist Tendencies

Western Province has stood apart from the rest of Zambia since the British South African Company first turned it into a semi-autonomous protectorate under the authority of the Lozi Litunga. When Zambia achieved independence, this unique status was recognized in the Barotseland Agreement of 1964, which granted what was then called Barotseland unique self-governance rights (Taylor, 2006; Englebert, 2005; Sardanis, 2014).

Starting in the early 1990s, however, tensions over this semi-autonomy increased markedly. During 1994, 3,000 Lozi formed a temporary army to defend their leader from the national government, a move deemed “treasonable by the government” (Englebert, 2005, p. 36). Police raids later led to the confiscation of rocket launchers, anti-aircraft guns, explosives, and hand grenades in the province. In 1995 the national government passed the Land Act, stripping the Litunga of control over the allocation of public lands, one of the Litunga’s core prerogatives (Litunga literally means “land” (Englebert, 2005, p. 42)). Tensions continued to escalate over the next decade, culminating in violent confrontations between police and protestors who were demanding the secession of Western Province in January of 2011 (Sardanis, 2014, p.220) and the establishment of the Barotse Freedom Movement and the Movement for the Restoration of Barotseland.
L Calculation of National Integration

National Integration for a Province $p \in P$ and a set of network communities $c \in C$ is

$$\text{Integration}_p = 1 - \sum_{c \in C} \left[ (\text{Share of users in community } c \text{ who reside in } p) \times (\text{Share of users who live in } p \text{ who are in } c) \right]$$

where the first term can be thought of as a measure of integration (increasing in the diversity of $c$), and the second term a weighting parameter (increasing in the share of residents in $p$). More intuitively, $\text{Integration}_p$ is the probability that if one were to select one resident of Province $p$ and another random member of that resident’s network community, they would come from different network communities.

\[\text{Residency is determined based on the centroid of each user’s Second-Order Theissen Polygon.}\]
M Computation of Weighted Fragmentation Measures

Available data can only geo-reference cell-phone subscribers to specific regions, not points (see Appendix F). As a result, situations arise where subscribers are assigned to regions that are not completely contained within an electoral district. When this occurs, users are assumed to be distributed in proportion to the area of the assigned region that falls within each district and each district’s population density. This gives rise to the following revised formulation of Equation 1 where, for a region to which subscribers are assigned $t \in T$ and electoral districts $d \in D$, let the share of subscribers in $t$ who reside in $d$ be defined as:

$$s(t,d^*) = \frac{\text{area}(t,d^*) \times \text{pop}_\text{density}(d^*)}{\sum_{d \in D} \text{area}(t,d) \times \text{pop}_\text{density}(d)}$$

(7)

where $\text{area}(t,d^*)$ is the area of overlap between $t$ and $d^*$ and $\text{pop}_\text{density}(d^*)$ is the population density of district $d^*$. As discussed in Appendix F, however, geo-referencing is actually done with respect to Wards, a lower administrative level than Constituencies, so weights are actually calculated at the Ward level with information on Ward population density and areas of intersection (i.e. $d$ is an index of Wards), then aggregated up to the Constituency level, improving precision. Wards are nested within Constituencies, so this strategy is strictly dominant. A map of Ward and Constituency boundaries is presented in Figure 7.

These shares can then be used as weights to compute a weighted version of Equation 1:

$$NF_d = 1 - \sum_{c \in C} \left( \frac{\sum_{t \in T} s(t,d)n_{c,t}}{\sum_{t \in T} s(t,d)n_t} \right)^2$$

(8)

where $n_t$ is the number of subscribers in region $t$ and $n_{c,d}$ is the number of subscribers in community $c$ in region $t$. 
N Robustness to Dropping Province Fixed-Effects
Figure 8: ELF and network fragmentation, No Province Fixed Effects

Network Fragmentation and ELF
No Province Fixed Effects

All-Days, > 2

Logged Resolution Parameter

Weekend-Only, > 2

Logged Resolution Parameter

All-Days, > 6

Logged Resolution Parameter

Weekend-Only, > 6

Logged Resolution Parameter
O Robustness of Rural-Sample Results

Figure 9

Rural Network Fragmentation and ELF
Population At Least 85% Rural

Weekend-Only, > 2

Weekend-Only, > 6

Includes 75 Constituencies with >85% rural populations.

Figure [11] presents a scatterplot of ELF and network fragmentation after controlling for the same factors as used in Figure 3 using a residual-regression framework. The left panel of Figure [11] shows the relationship when network fragmentation is computed using the Significance-optimizing resolution parameter, while the right panel presents the relationship using network fragmentation using the resolution parameters identified in Section 7. Consistent with previous results, the plots shown a consistently positive relationship between ELF and network fragmentation for the rural sample, and importantly one not driven by any outliers.

12 Both ELF and network fragmentation are regressed against all controls. Residuals from the two regressions are then plotted against one another.

13 The logged resolution closest to the average of logged resolutions shown in Table 8 is used when the average value lies between calculated values.
Figure 10

Rural Network Fragmentation and ELF
Limited Controls

All-Days, > 2

All-Days, > 6

Weekend-Only, > 2

Weekend-Only, > 6

Includes 104 Constituencies with >70% rural populations.

Figure 11: ELF and Network Fractionalization, Rural Sample

Significance-Optimizing Resolution Parameter

Substantively-Chosen Resolution Parameter

95% Confidence Intervals Plotted
Standard Errors clustered at Province level
Includes 104 Constituencies at least 70% rural populations
Figure 12: ELF and Network Fragmentation, Rural Sample, No Province Fixed Effects

Network Fragmentation and ELF
70% Rural, No Province Fixed Effects

All-Days, > 2

All-Days, > 6

Weekend-Only, > 2

Weekend-Only, > 6

Logged Resolution Parameter

xxvii
P Selection of Public Goods and Voter Knowledge

Voter knowledge and public goods were chosen because of the clear predictions made about their empirical behavior by network theory. By contrast, while network theory makes clear predictions about the relationship between network fragmentation and capacity to organize protests or coordinate voters, it does not make clear predictions about when these behaviors should be observed. For example, to the degree to which protests can be thought of as a sanction against poorly performing politicians, they constitute an off-the-equilibrium-path outcome. Politicians should respond to greater protest capacity by improving performance, diminishing the likelihood of observing protests. Off-the-equilibrium-path trembles certainly occur, but it is not clear whether those trembles are more likely in places with higher protest capacity. Similarly, while situations exist in which voter coordination is clearly important for electoral effectiveness, measuring coordination is non-trivial. First, voter coordination is a behavior that takes place among voters with similar preferences; a community that divides its vote over many candidates may be failing to coordinate, but it also may just have heterogeneous preferences over candidates. As a result, coordination is difficult to measure absent clear measures of underlying voter preferences. But second, coordination is not always an optimal strategy. Coordination only matters in a first-past-the-post electoral system like Zambia when (a) one of the preferred candidates for a pool of voters would win if said voters coordinated, but (b) both of those preferred candidates candidate would lose if they failed to coordinate. Provided coordination is costly, coordination should only occur up to this threshold (where one of the preferred candidates receives 50% of votes, plus perhaps some margin for risk aversion). Thus without clear knowledge about the underlying candidate preferences of all voters, knowing whether a community is in a situation in which coordination is called for is not empirically feasible. [Rojo and Wibbels] (2014) offers an alternate and somewhat dissenting view on the strategic rational for voter coordination that also does not translate well to the higher geographic units for which vote totals are available in Zambia.

Q Political Knowledge Questions

- “Can you tell me the name of: Your Member of Parliament?”
  - **Responses Coded as Knowledgeable**: Know but can’t remember, Incorrect guss, Correct name, Refused to Answer.
  - **Responses Coded as Note Knowledgeable**: Don’t Know
- “How much time does your Member of Parliament spend in this constituency?”
  - **Responses Coded as Knowledgeable**: Never, At Least Once a Year, At Least Once
a Month, At Least Weekly, She/ He is Here Almost All the Time, Refused to Answer

- **Responses Coded as Note Knowledgeable**: Don’t Know

- “How much of the time do think the following try their best to listen to what people like you have to say: Members of Parliament?”
  - **Responses Coded as Knowledgeable**: Never, Only Sometimes, Often, Always, Refused to Answer
  - **Responses Coded as Note Knowledgeable**: Don’t Know

- “Do you approve or disapprove of the way the following people have performed their jobs over the past twelve months, or haven’t you heard enough about them to say: Your Member of Parliament (MP)?”
  - **Responses Coded as Knowledgeable**: Strongly Disapprove, Disapprove, Approve, Strongly Approve, Refused to Answer
  - **Responses Coded as Note Knowledgeable**: Don’t Know

- “How well or badly would you say the current government is handling the following matters, or haven’t you heard enough about them to say: [policy area]”
  - **Policy Areas**:
    * Managing the economy
    * Creating jobs
    * Keeping prices stable
    * Narrowing income gaps
    * Reducing crime
    * Improving basic health services
    * Addressing educational needs
    * Delivering household water
    * Ensuring enough to eat
    * Fighting corruption
    * Combating HIV/AIDS
  - **Responses**:
    * **Responses Coded as Knowledgeable**: Very Badly, Fairly Badly, Fairly Well, Very Well, Refused to Answer
    * **Responses Coded as Note Knowledgeable**: Don’t Know

---

14 Note that in the survey context, “current government” clearly refers to National Government.
R Computing a Public Goods Index

Public Goods Data

Data on delivery of these services comes from a 10% sample of micro-data from the 2000 and 2010 Zambian National Censuses, which includes data on four essential public services:

- **Electrification**: Share of individuals living in a household that reports “electricity” as their main source of lighting.\(^\text{15}\)
- **Protected Water Supply**: Share of individuals living in a household that reports their main source of water is either “Piped inside”, “Piped outside within plot”, “Communal tap”, “Protected well”, or “Protected borehole”.\(^\text{16}\)
- **Enrollment**: Share of children aged 5 to 15 (inclusive) reported as currently attending school.
- **Infant mortality**: Share of children born in the past year who are now deceased.\(^\text{17}\)

Note that while this data is provided at the level of households, as a result of confidentiality concerns, the only geographic identifier associated with households is their Constituency and whether they are urban or rural. As noted above, however, this is not especially problematic. The National Government is solely responsible for schooling and most aspects of the health system, and it plays an extremely active role in both electrification and water provision. As such, our primary interest is in outcomes at the level of national legislature electoral districts (Constituencies).

The census data used here has four major advantages over other measures of public goods. First, unlike one-off surveys of government facilities, the repeat cross-section of census data allows for measurement of changes in public service provision, rather than the level of infrastructure present. Government facilities – like schools, health centers and roads – represent a stock of resources accumulated over decades of investment, and therefore represent the aggregate outcome of constantly changing political processes.\(^\text{18,19}\) Second, unlike somewhat richer

\(^{15}\) In the 2000 census, this number was 19%, while the share reporting “Paraffin” was 51% and “Candle” was 16%.

\(^{16}\) In the 2000 census, this number was 49%. An additional 28% reported their primary source was an unprotected well, and 20% reported relying on “River, dam, or stream”.

\(^{17}\) Note that because surveyors ask about the survival rate of any live births occurring in the past year, all relevant births will have occurred within the last year. As a result, this measure is distinct from the number of children who survive to age 1. The average Constituency reported a mortality rate of 5.1% in 2000.

\(^{18}\) This problem would be less severe if surveys included the date of facility constructed. Unfortunately, data sources this author has been able to locate – like the 2005 Ministry of Education survey of Schools in Zambia, the 2006 Southern African Human-development Information Management (SAHIMS) survey of Zambian Health Facilities, or the Global Roads Open Access Data Set (gROADS v.1) estimates of road density – do not include this information.

\(^{19}\) Substantiating the idea that measures of “stock” represent the aggregation of ever changing political processes...
household surveys like the Demographic Health Surveys (DHS), the census is fully representative at all levels, allowing for an analysis of all Constituencies. Third, unlike data on budget allocations – which has been used by some other authors in Zambia (e.g. Gibson and Hoffman [2013]) – census data provides a measure of outcomes, rather than inputs. Given that the mapping from inputs to outputs depends on the effort and honesty of government officials, and given that our normative interest is in the quality of actual human development, measurement of outcomes is preferable. Finally, and perhaps most importantly, the range of measures included in this data make it possible to address the “basket of goods” problem by looking at several outcomes.

Motivation for Index Use

The use of an index is motivated by the observation that politicians are often responsible for a wide portfolio of goods and services, and failing to take this into account by focusing only on a single public good can often lead to erroneous conclusions (the “basket of goods” problem [Kramon and Posner, 2013]). For example, consider a politician who chooses to prioritize schooling over health investments. If one measures only schooling or health outcomes, that politician will appear effective by one measure and ineffective by another, while the truth likely lies somewhere in the middle.

Computing a Public Goods Index

The four public good measures available in the census are aggregated into a public goods index, computed as the first component of a Principle Component Analysis (PCA) of the normalized values of these measures (mean zero, standard deviation of one). The aim of this index is to find a measure of the common process – traditionally thought of as government competence or responsiveness.

Prior to execution of the PCA, however, one additional adjustment is made to public goods measures. Zambia, like many countries in the region, experienced a dramatic increase in urbanization during the 2000s. As a result of this movement, many more people had access to electricity and protected water sources in 2010 than in 2000 due solely to where they chose to live. Since this change was in no way due to improvements in government performance, an ideal measure of public goods improvement should correct for these changes. To do so, improvement where investments in different public goods may have taken place at different types under different pressures, measures of levels of public goods are not only not well correlated with one another, but in many cases are actually negatively correlated.
in public goods for Constituency $i$ is computed here as follows, where $PG$ is the level of access to a public good:

$$
\Delta PG_i = (PG_{2010,i,rural} - PG_{2000,i,rural}) \times (\text{share of population rural}_{2010,i}) \\
+ (PG_{2010,i,urban} - PG_{2000,i,urban}) \times (\text{share of population urban}_{2010,i})
$$

(9)

This correction significantly improves the correlation between improvement scores for different public goods, increasing the likelihood that the measure is capturing a common process like government responsiveness. This is shown in Tables 1 and 2, which show the correlation between measure of public goods when one considers the simple change from 2000 to 2010 in each Constituency (Table 1) and when one corrects for urbanization within the Constituency using Equation 9 (Table 2). (The sign on “infant mortality” has been flipped so positive values represent improvements in infant mortality rates.) As the two tables show, adjustment for urbanization leads to substantial improvements in inter-public-good-measure correlations, suggesting a substantial improvement in the degree to which these measure a common process.

Table 1: Raw Inter-Public Good Measure Correlations

<table>
<thead>
<tr>
<th>Variables</th>
<th>Water</th>
<th>Electricity</th>
<th>Enrollment</th>
<th>Infant Mort.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electricity</td>
<td>-0.12</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enrollment</td>
<td>0.44</td>
<td>0.01</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Infant Mort.</td>
<td>0.02</td>
<td>0.05</td>
<td>0.06</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 2: Corrected Inter-Public Good Measure Correlations

<table>
<thead>
<tr>
<th>Variables</th>
<th>Water</th>
<th>Electricity</th>
<th>Enrollment</th>
<th>Infant Mort.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electricity</td>
<td>-0.03</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enrollment</td>
<td>0.46</td>
<td>0.05</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Infant Mort.</td>
<td>0.04</td>
<td>0.15</td>
<td>0.10</td>
<td>1.00</td>
</tr>
</tbody>
</table>

After normalization, measures of change in public goods are aggregated into a single measure using a PCA index. As shown in Table 3 below, the resulting measures all have positive first component loads, again suggesting the index is indeed capturing a common latent process.20

20It is important to emphasize that the creation of an index is not a panacea for the “basket of goods” problem, however. Calculation of an index requires the researcher to assign weights to the relative importances of changes in different public services, weights which may or may not correspond to true social-welfare values. These weightings
Table 3: Public Goods PCA Loads

<table>
<thead>
<tr>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
</tr>
<tr>
<td>Electricity</td>
</tr>
<tr>
<td>Enrollment</td>
</tr>
<tr>
<td>Infant</td>
</tr>
</tbody>
</table>

S Public Goods, Voter Knowledge, and Network Fragmentation

Because of large variation in the standard errors associated with estimates generated in Figure 5, it is not possible to represent all estimates in a single plot in a manner that properly conveys all relevant information. With that in mind, this Appendix presents the estimates from Figure 5 normalized by the standard deviation of network fragmentation rather than the standard error of the estimate (so units are in changes in voter knowledge / public goods associated with a one standard deviation change in network fragmentation) with various y-axis scales. These figures are all plotting the same estimates, they are simply plotting different sub-samples with different associated ranges of the y-axis.

are often only implicit, but are always present. In this case, the choice to use normalized values of the different public service measures amounts to assigning equal importance to the social welfare value on a standard deviation change in each measure, a strategy analogous to that employed by [Anderson 2008]. This equal-weighting strategy has the advantage of minimizing researcher discretion to ameliorate concerns about data-mining [Roodman 2005], but it is important to recognize that any weighting decisions have important normative implications, and equal weighting is neither a benign nor assumption free approach [Knack, Rogers and Eubank 2011].
Network Fragmentation and Public Goods
By Resolution Parameter and Filter

Network Fragmentation and Political Knowledge
By Resolution Parameter and Filter

Figure 13: Estimates for $\gamma \in [-16, -8]$
Network Fragmentation and Public Goods
By Resolution Parameter and Filter

Network Fragmentation and Political Knowledge
By Resolution Parameter and Filter

Figure 14: Estimates for $\gamma \in [-16, -6]$
Network Fragmentation and Public Goods
By Resolution Parameter and Filter

All-Days, > 2

Weekend-Only, > 2

All-Days, > 6

Weekend-Only, > 6

Network Fragmentation and Political Knowledge
By Resolution Parameter and Filter

All-Days, > 2

Weekend-Only, > 2

All-Days, > 6

Weekend-Only, > 6

Figure 15: Estimates for $\gamma \in [-16, -4]$
Figure 16: Estimates for $\gamma \in [-16, 0]$

## Results By Resolution Selection Criterion

The resolutions that are most correlated with voter knowledge and public goods are quite different from those identified using the atheoretical criterion. Table 4 presents descriptive statistics for communities under the atheoretic criterion (“Atheoretic”) and under the resolution parameters identified in this section (“Theory”). In nearly all cases the size of the average community is higher under the newly identified parameter. Even in the “All-Days > 2” network, the median community size increases substantially. (At the resolution selected by atheoretical criterion, the community detection algorithm identified a few very large networks and lots of very
small ones. At the new parameter, communities are larger on average, but no super-communities exist.

This result comports well with political intuition. Communities at the resolution selected by atheoretic criterion may be clearly delineated, but their median size of 24-42 people seems relatively small to be politically effective. The resolution that most correlates with voter knowledge and public goods generates communities with median sizes of 184-1,406 people – much closer to the scale of a political protest or rally – suggesting a “political relevant” social scale.
Table 4: Community Sizes

<table>
<thead>
<tr>
<th></th>
<th>Atheoretic Mean</th>
<th>Atheoretic Median</th>
<th>Atheoretic Max</th>
<th>Theory Mean</th>
<th>Theory Median</th>
<th>Theory Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-Days, &gt;2</td>
<td>8,058</td>
<td>34</td>
<td>88,724</td>
<td>759</td>
<td>258</td>
<td>14,610</td>
</tr>
<tr>
<td>All-Days, &gt;6</td>
<td>50</td>
<td>42</td>
<td>507</td>
<td>328</td>
<td>184</td>
<td>5,471</td>
</tr>
<tr>
<td>Weekend-Only, &gt;2</td>
<td>45</td>
<td>38</td>
<td>401</td>
<td>896</td>
<td>265</td>
<td>19,677</td>
</tr>
<tr>
<td>Weekend-Only, &gt;6</td>
<td>26</td>
<td>24</td>
<td>195</td>
<td>882,060</td>
<td>1,406</td>
<td>1,966,897</td>
</tr>
</tbody>
</table>
U Alternate Cleavages

Tables 5 presents summary statistics for the distribution of individuals across religions in the 2010 Zambian Census Microdata 10% public sample. Categories for religious identification come directly from the Zambian census.

Table 5: Distribution of Religious Identities

<table>
<thead>
<tr>
<th>Item</th>
<th>Number</th>
<th>Per cent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catholic</td>
<td>253614.00</td>
<td>20.29</td>
</tr>
<tr>
<td>Protestant</td>
<td>940328.00</td>
<td>75.25</td>
</tr>
<tr>
<td>Muslim</td>
<td>5,776.00</td>
<td>0.46</td>
</tr>
<tr>
<td>Hindu</td>
<td>416.00</td>
<td>0.03</td>
</tr>
<tr>
<td>Buddist</td>
<td>905.00</td>
<td>0.07</td>
</tr>
<tr>
<td>Bahai faith</td>
<td>327.00</td>
<td>0.03</td>
</tr>
<tr>
<td>Other</td>
<td>25,487.00</td>
<td>2.04</td>
</tr>
<tr>
<td>None</td>
<td>22,832.00</td>
<td>1.83</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1.25e+06</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Note that results are similar when calculating fragmentation using only the categories of Protestant and Non-Protestant.

Table 6 presents summary statistics for the distribution of individuals across religions in the 2010 Zambian Census Microdata 10% public sample. Categories for employment industries were constructed using the much more granular industry strings in census variable p35. Aggregation was necessary as the variable included over 235 distinct string responses, many referring to very similar activities. Aggregation was undertaken using standard sector definitions with the exception of mining, which in the Zambian context appears worthy of it’s own designation.

Coding rules are as follows:

- **Farming/Fishing/Forestry**: Low-value cultivation of primary goods.
  
  Examples:
  - Farming / Agriculture
  - Logging
  - Fishing

- **Manufacturing**: Activities with the primary aim of creating a physical object.
  
  Examples:
  - Any activity described as manufacturing
  - Construction
- Spinning and weaving of textiles
- Sawmilling

**Services:** *Activities where the output of productive activity is not a physical object.*

Examples:
- Administrative services
- Retail sales
- Legal and financial services
- Teaching
- Healthcare services

**Mining:** *Mining or quarrying.*

### Table 6: Distribution of Industries of Employment

<table>
<thead>
<tr>
<th>Item</th>
<th>Number</th>
<th>Per cent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturing</td>
<td>21,663.00</td>
<td>6.19</td>
</tr>
<tr>
<td>Farming</td>
<td>242731.00</td>
<td>69.41</td>
</tr>
<tr>
<td>Service</td>
<td>78,507.00</td>
<td>22.45</td>
</tr>
<tr>
<td>Mining</td>
<td>6,799.00</td>
<td>1.94</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>349700.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

V Additional Acknowledgements

In addition to parties cited in the body of the paper, the author is also indebted to numerous open source software contributors, including the authors of *iGraph* [Csardi and Nepusz (2006)], *pandas*, and *matplotlib*.

This project would also not have been possible with the support of the team behind [Anatella](https://anatella.com) ETL.
References


Motahari, Sara, Sandeep Appala, Ole J Mengshoel, Luca Zoia, Phyllis Reuther and Jay Shah. 2012. The Impact of Social Affinity on Phone Calling Patterns: Categorizing Social Ties from Call Data Records - Semantic Scholar. In The Sixth SNA-KDD Workshop.


Wiese, Jason, Jun-Ki Min, Jason I Hong and John Zimmerman. 2015. You Never Call, You Never Write: Call and SMS Logs Do Not Always Indicate Tie Strength. Call and SMS Logs Do Not Always Indicate Tie Strength New York, New York, USA: ACM.
