Appendix to “Social Networks and the Political Salience of Ethnicity”

Nicholas Eubank

Date: August 3, 2018

Appendix Contents

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 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.[1] 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

[1]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"[2] (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% 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.[3] 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)

[3]Individuals 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.[4] 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.[5] 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.

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

[5]Though 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 σ 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 σ_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 σ and summing across all subscribers $i, j \in V$, this intuition can be captured with the following objective function:

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

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 σ_i is equal to subscriber j's community σ_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:

$$\mathcal{H}(\sigma, \gamma) = -\sum_{ij \in V} [A_{ij} - \gamma]\delta(\sigma_i, \sigma_j) \tag{2}$$

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

Note that as γ 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 γ 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 γ for each member of the existing community with whom it is not connected. As a result, the higher the value of γ, 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, γ is often referred to as the *resolution parameter* of a CPM function.

Because γ determines the size and density of network communities, the choice of γ is directly tied to the problem of scale described in the introduction. The choice of a large value of γ (high resolution) will generate very small, very densely connected communities (like a group of households within a village), while a low value of γ will give rise to large but more loosely connected communities (more like a group of villages within a region).[6]

[6]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 (Fortunato and Barthelemy, 2007; Traag, Van Dooren and Nesterov, 2011).

E Optimization of the Constant Potts Model

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.

F 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.[7] 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:

$$T(a) \equiv \{p : \|a - p\| < \|a' - p\| \quad \forall a' \in A \quad where \quad a' \neq a\} \tag{3}$$

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:

$$\begin{aligned} T(a) \equiv \{p : \quad & \|a - p\| < 35km \,\& \\ & \|a - p\| < \|a' - p\| \quad \forall a' \in A \quad where \quad a' \neq a\} \end{aligned} \tag{4}$$

An illustration of this type of Thiessen polygon is given in Figure 1 (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

[7]When a subscriber places a call, the subscriber's phone chooses a cell-tower through which to route the call. Cell-towers are chosen to maximize call quality, taking into account factors like signal strength and the number of other callers using each tower. Because these factors vary over time, calls placed from the same position may be routed through different towers at different times. However, signal strength decreases quickly with distance – even in a vacuum, radio strength decreases as the inverse square of distance – so calls are most likely to be routed through the nearest tower.

Thiessen polygon associated with a primary tower a_1 and a secondary tower a_2 is defined as:

$$T(a_1, a_2) \equiv \{p : \quad \begin{aligned} & \|a_1 - p\| < 35km \quad \& \\ & \|a_1 - p\| < \|a' - p\| \quad \forall a' \in A \quad where \quad a' \neq a_1 \quad \& \\ & \|a_2 - p\| < \|a'' - p\| \quad \forall a'' \in A \quad where \quad a'' \neq a_1, a_2\} \end{aligned} \tag{5}$$

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 Theissen 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.

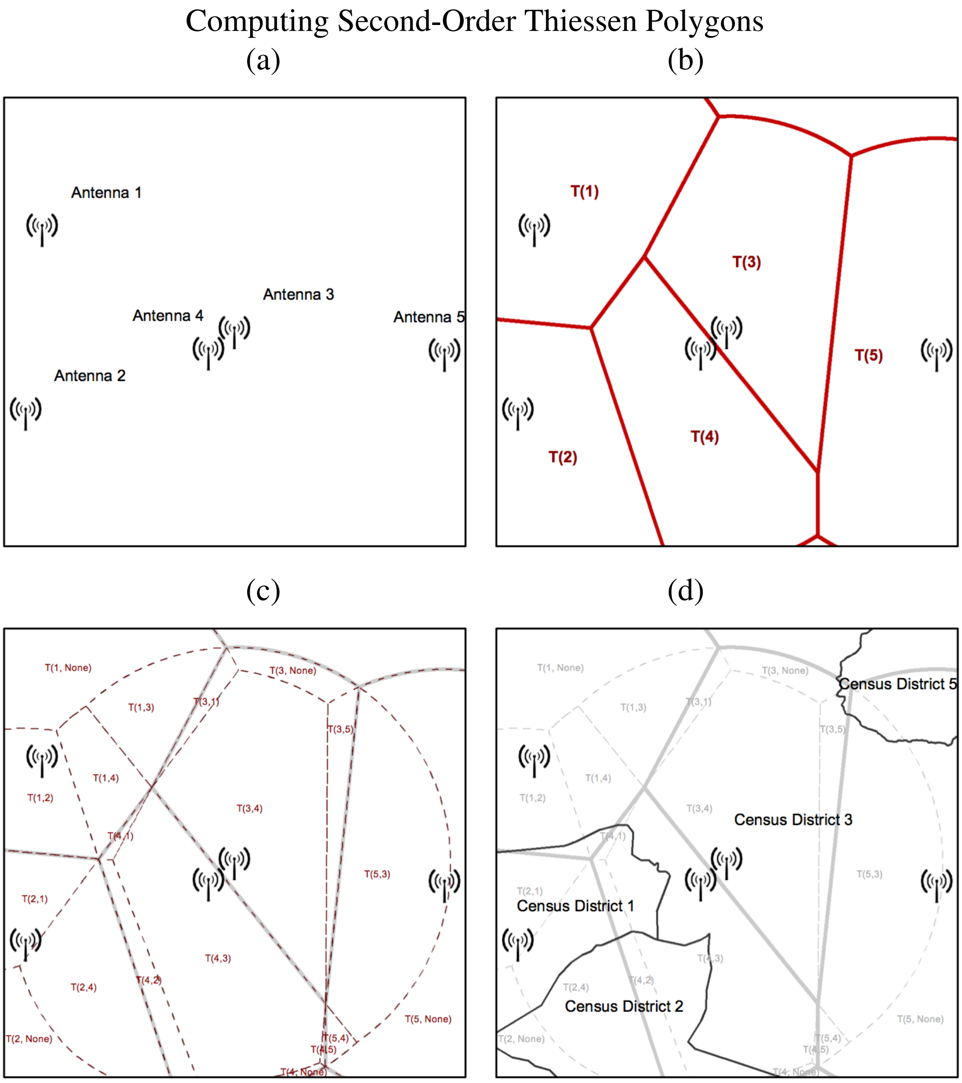


Figure 1: Geo-referencing proceeds in four steps. In **(a)**, cell-phone antenna towers are geo-referenced. In **(b)**, Zambia is partitioned into regions based on the identity of the *closest* cell-phone antenna tower. At all points within $T(3)$, for example, Antenna 3 is the closest cell-phone tower. 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 Theissen Polygons. Finally, in **(d)** these Second-Order Theissen 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" (Sardanis, 2014, p.240).

H Community Size Histogram

Figure 2 plots histograms of the network community sizes summarized in Section 4.

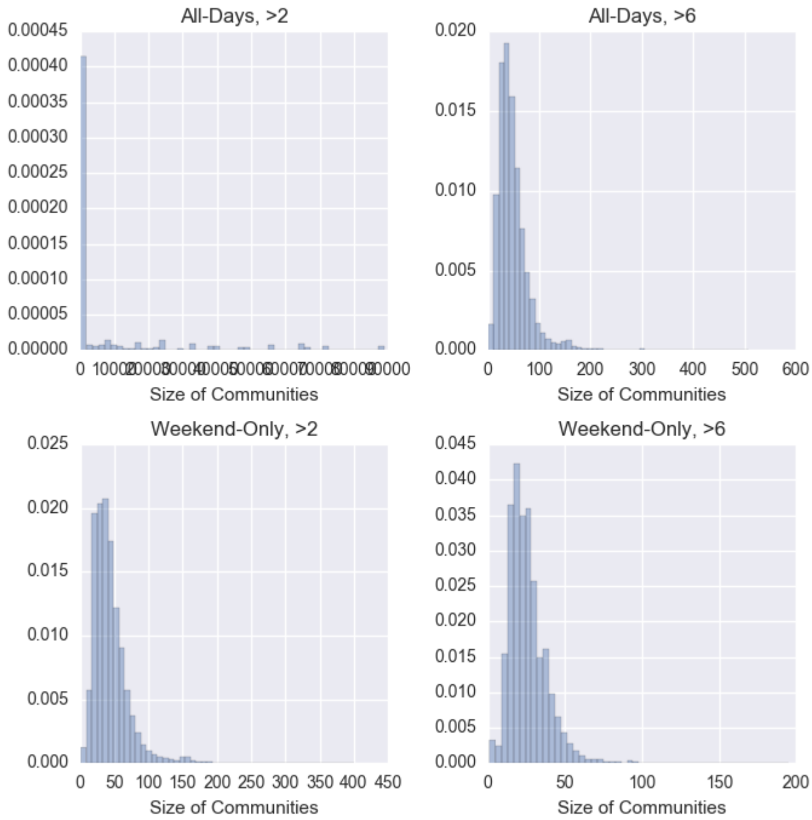


Figure 2: Distribution of network community sizes. Each plot shows a histogram of network community sizes under different network specification.

I Spatial Distribution of Communities

Figure 3: Spatial Distribution of 8 Random Communities, Most Inclusive Network

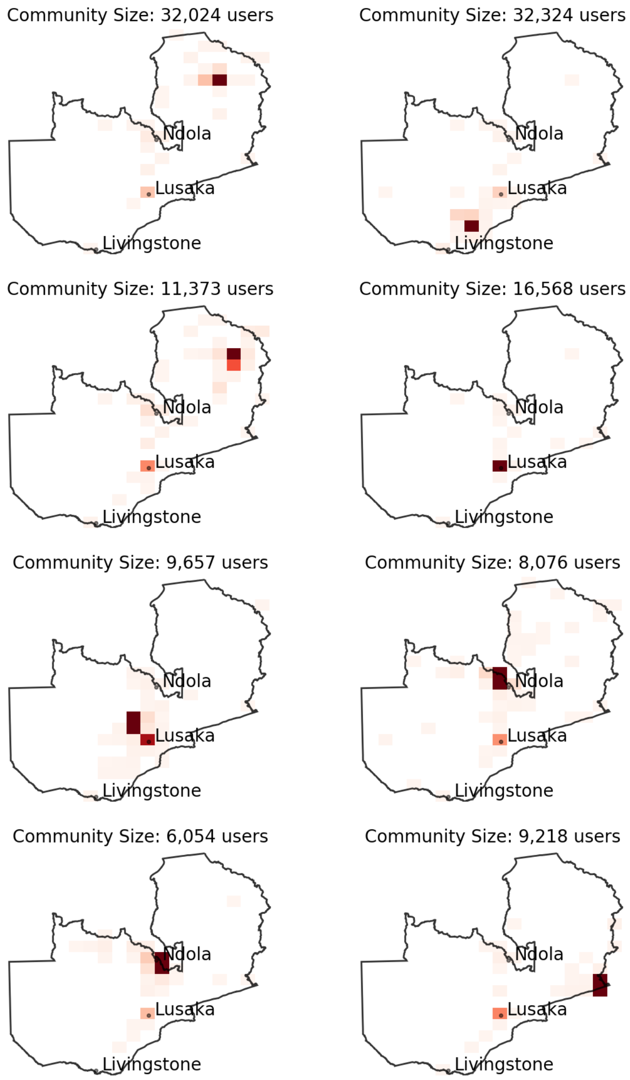


Figure 4: Spatial Distribution of 8 Random Communities, All > 6

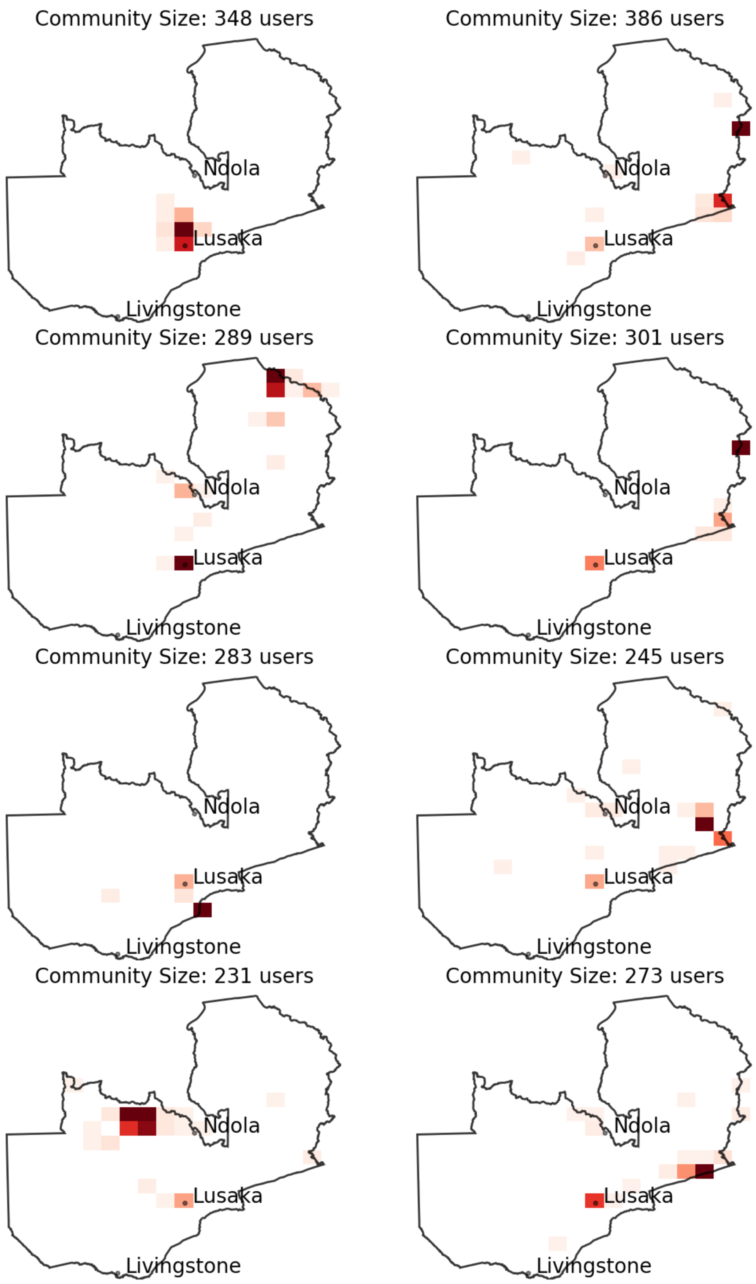
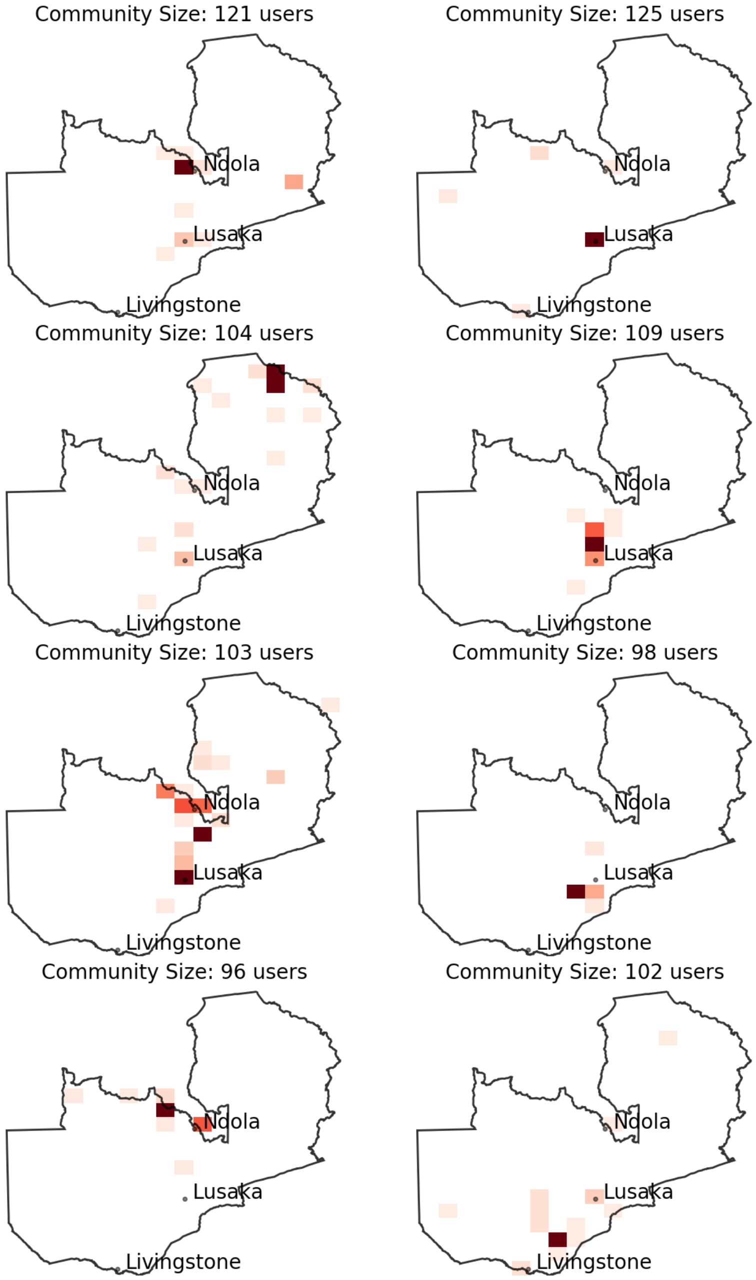


Figure 5: Spatial Distribution of 8 Random Communities, Weekends > 6



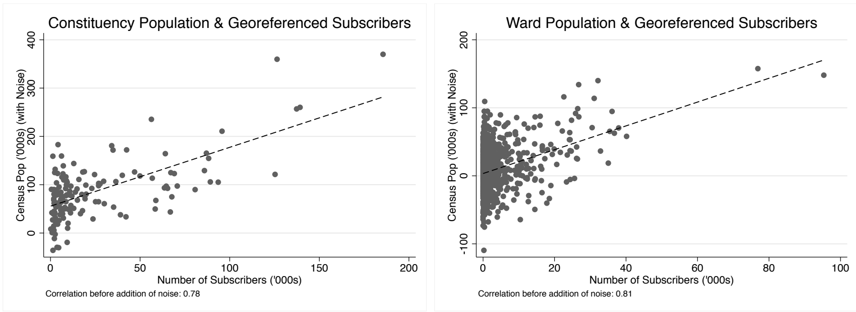


Figure 6: Scatter plots of census population and number of geo-referenced subscribers. Note that random noise has been added to census populations to protect commercially sensitive information. Actual correlations without noise presented below each figure.

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.[10]

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.

[10]The number of subscribers per district is commercially sensitive and without noise, census populations could be used to identify Constituencies.

K 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:[11]

$$Integration_p = 1 - \sum_{c \in C} [\quad (\text{Share of users in community } c \text{ who reside in } p) \qquad (6)$$
$$* (\text{Share of users who live in } p \text{ who are in } c)]$$

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, $Integration_p$ 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.

[11] 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:

$$s(t, d^*) \equiv \frac{area(t, d^*) * pop\_density(d^*)}{\sum_{d \in D} area(t, d) * pop\_density(d)} \tag{7}$$

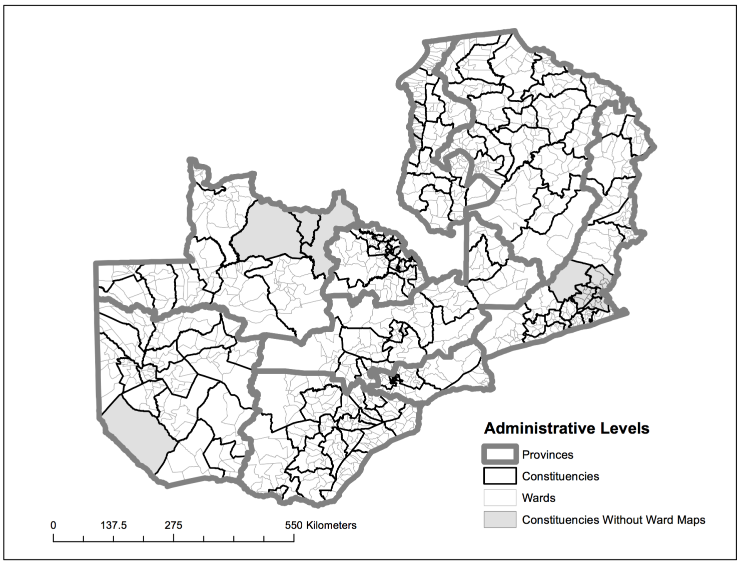
where $area(t, d^*)$ is the area of overlap between t and d^* and $pop\_density(d^*)$ is the population density of district d^*. As discussed in Appendix F, however, geo-referencing is actually done with respect to Wards, a lower administrative level than Constituencies, so weights are actually calculated 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:

$$NF_d = 1 - \sum_{c \in C} \left(\frac{\sum_{t \in T} s(t, d) n_{c,t}}{\sum_{t \in T} s(t, d) n_t} \right)^2 \tag{8}$$

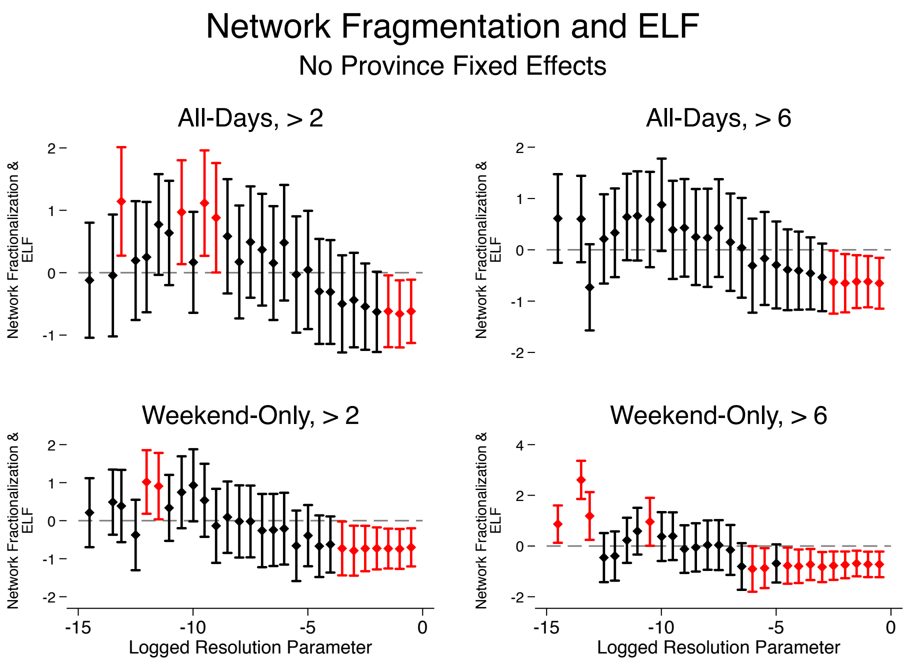
where n_t is the number of subscribers in region t and $n_{c,t}$ is the number of subscribers in community c in region t.

Figure 7: Administrative Boundaries



N Robustness to Dropping Province Fixed-Effects

Figure 8: ELF and network fragmentation, No Province Fixed Effects



O Robustness of Rural-Sample Results

Figure 9

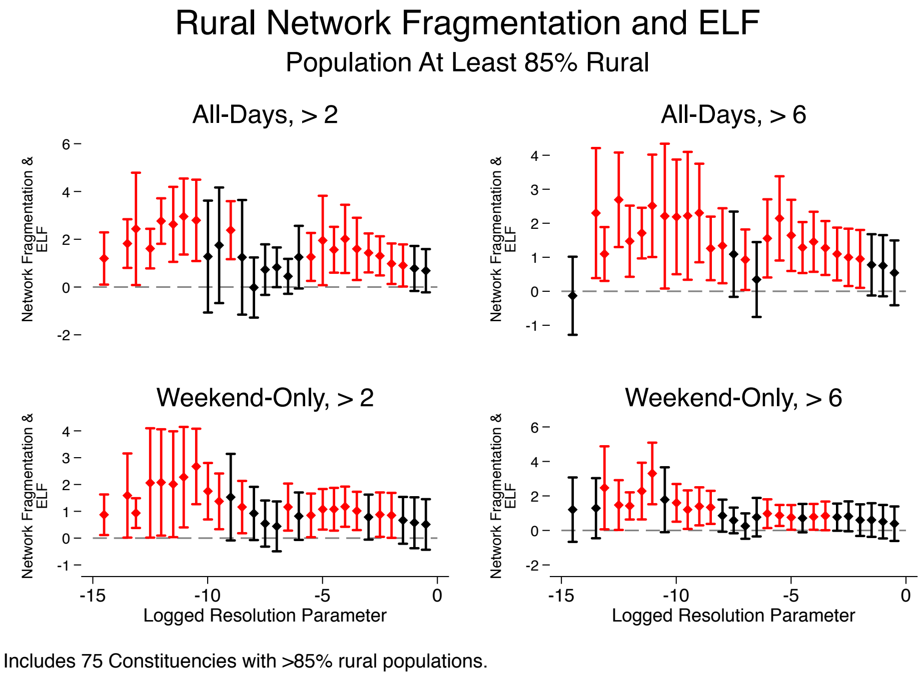


Figure 11 presents a scatterplot of ELF and network fragmentation after controlling for the same factors as used in Figure 3 using a residual-regression framework.[12] The left panel of Figure 11 shows the the relationship when network fragmentation is computed using the *Significance*-optimizing resolution parameter, while the right panel presents the relationship using network fragmentation using the resolution parameters identified in Section 7.[13] 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

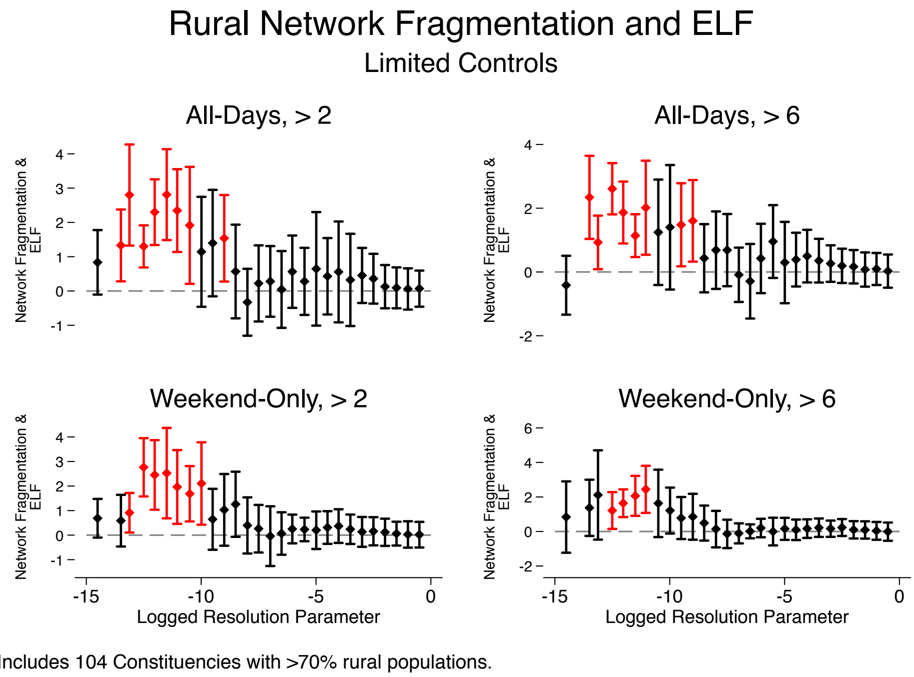


Figure 11: ELF and Network Fractionalization, Rural Sample

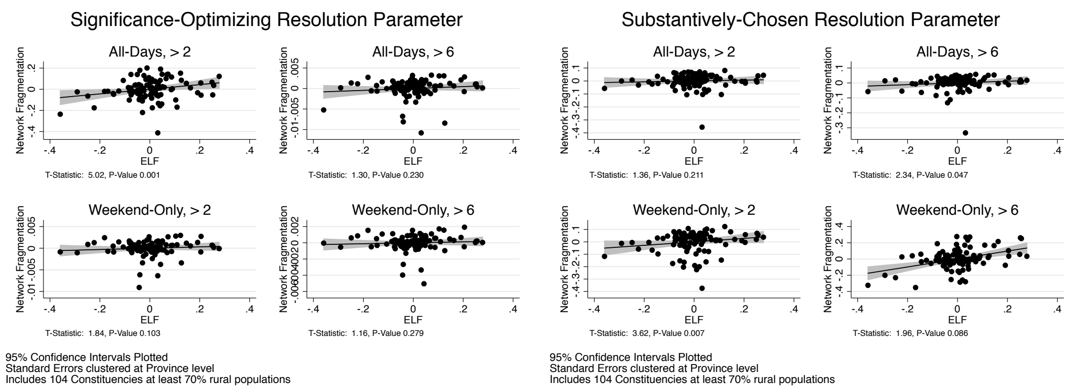
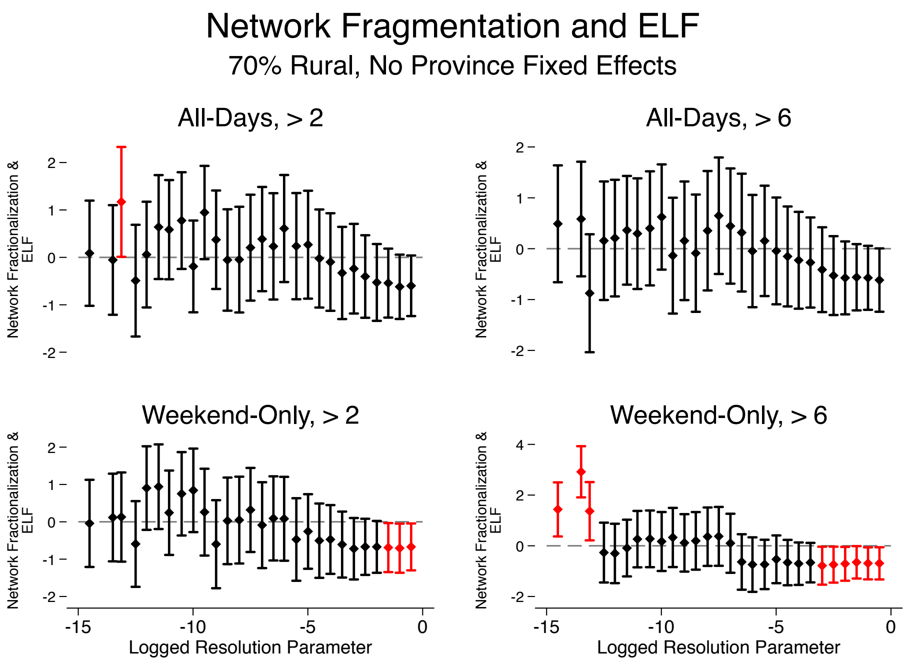


Figure 12: ELF and Network Fragmentation, Rural Sample, No Province Fixed Effects



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]"[14]
  - **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.[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".[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.[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.[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:

$$\begin{aligned}\Delta PG_i = \quad & (PG_{2010,i,rural} - \mathrm{PG}_{2000,i,rural}) * (\text{share of population rural}_{2010,i}) \\ & +(PG_{2010,i,urban} - PG_{2000,i,urban}) * (\text{share of population urban}_{2010,i})\end{aligned} \tag{9}$$

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

Variables	Water	Electricity	Enrollment	Infant Mort.
Water	1.00			
Electricity	-0.12	1.00		
Enrollment	0.44	0.01	1.00	
Infant Mort.	0.02	0.05	0.06	1.00

Table 2: Corrected Inter-Public Good Measure Correlations

Variables	Water	Electricity	Enrollment	Infant Mort.
Water	1.00			
Electricity	-0.03	1.00		
Enrollment	0.46	0.05	1.00	
Infant Mort.	0.04	0.15	0.10	1.00

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]

[20] It 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

	Score
Water	0.67
Electricity	0.11
Enrollment	0.70
Infant	0.23

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).

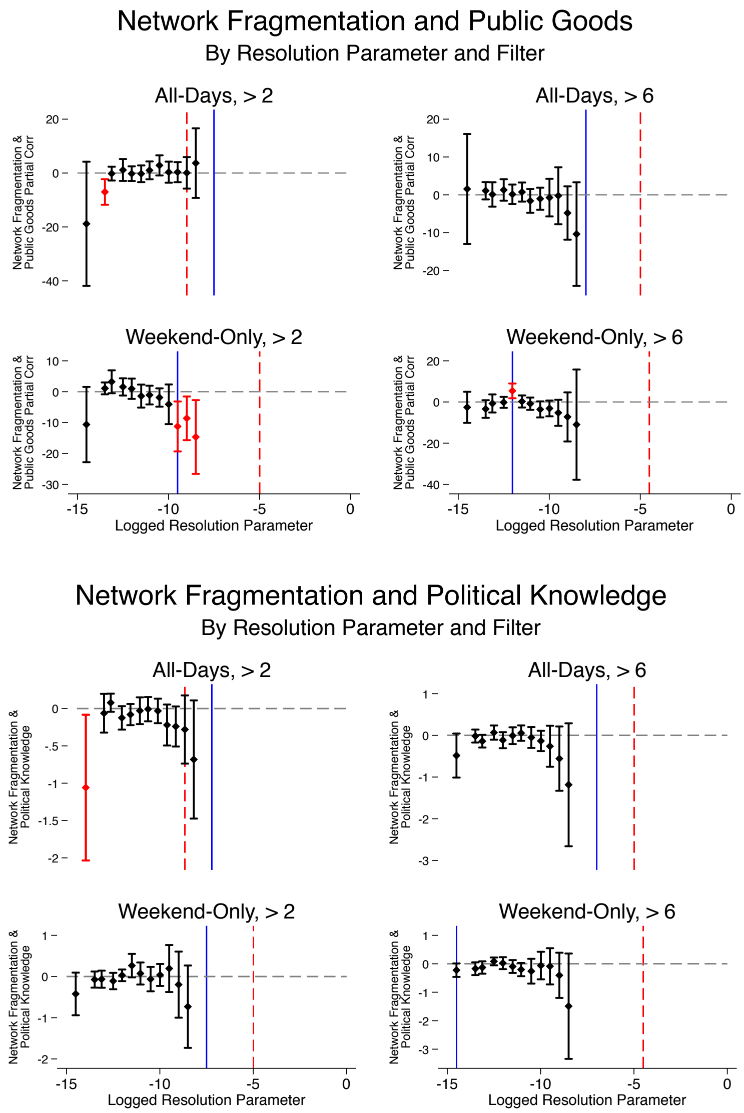
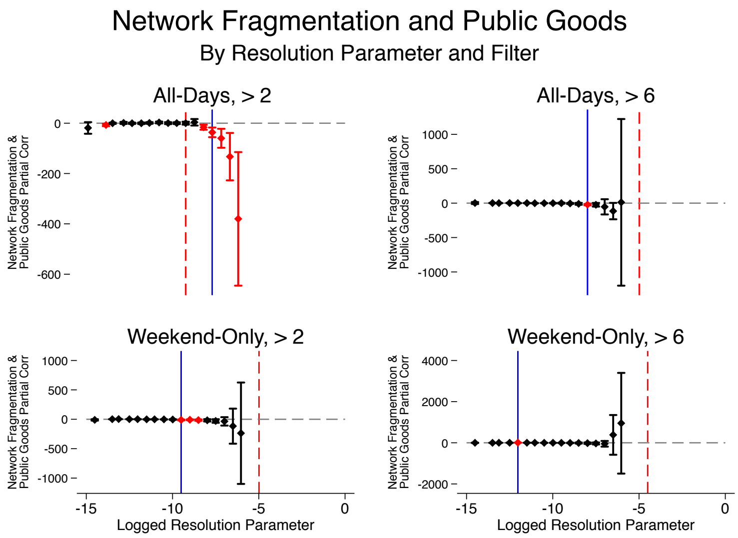


Figure 13: Estimates for $\gamma \in [-16, -8]$



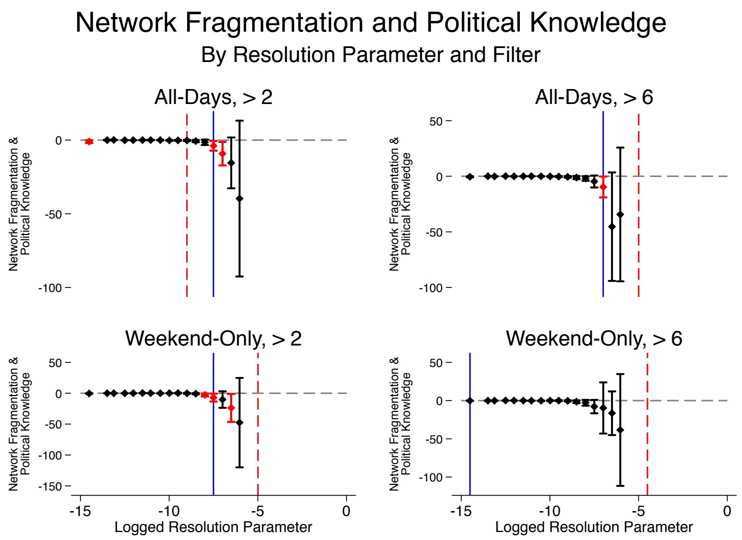
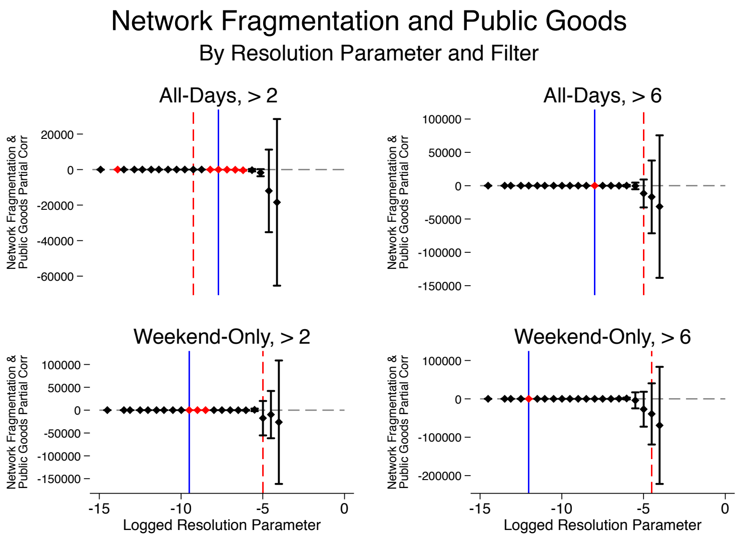


Figure 14: Estimates for $\gamma \in [-16, -6]$



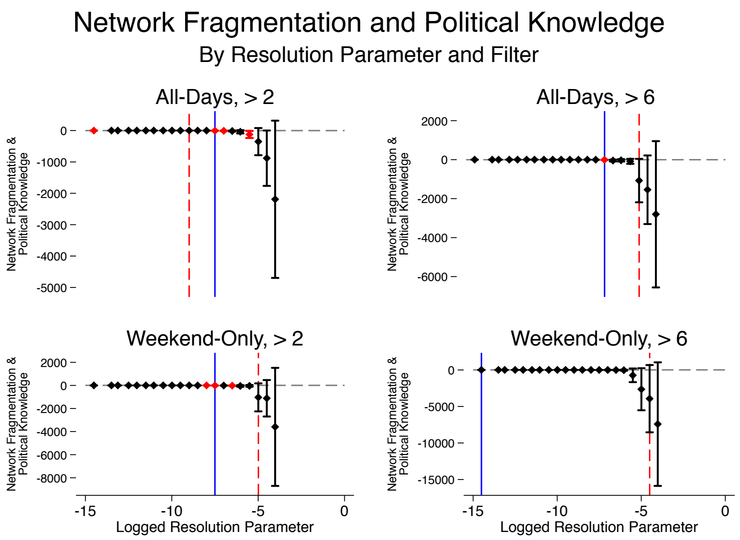


Figure 15: Estimates for $\gamma \in [-16, -4]$

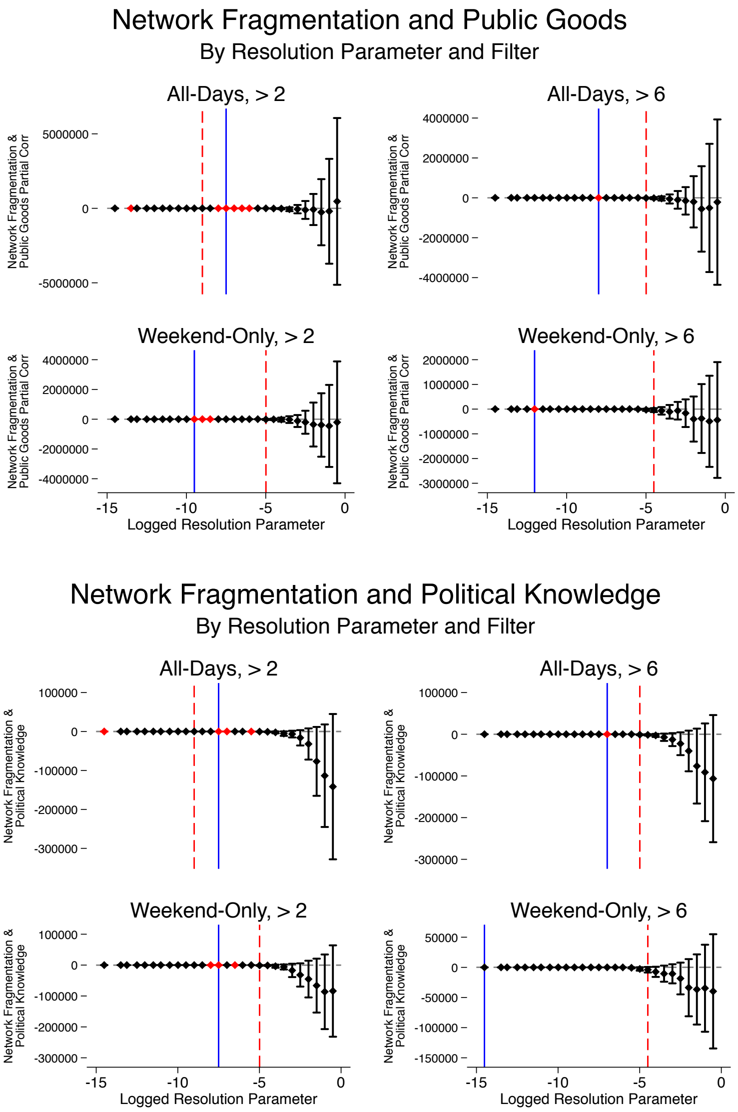


Figure 16: Estimates for $\gamma \in [-16, 0]$

T 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

	Atheoretic Mean	Atheoretic Median	Atheoretic Max	Theory Mean	Theory Median	Theory Max
All-Days, >2	8,058	34	88,724	759	258	14,610
All-Days, >6	50	42	507	328	184	5,471
Weekend-Only, >2	45	38	401	896	265	19,677
Weekend-Only, >6	26	24	195	882,060	1,406	1,966,897

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**

Item	Number	Per cent
Catholic	253614.00	20.29
Protestant	940328.00	75.25
Muslim	5,776.00	0.46
Hindu	416.00	0.03
Buddist	905.00	0.07
Bahai faith	327.00	0.03
Other	25,487.00	2.04
None	22,832.00	1.83
Total	1.25e+06	100.00

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**

Item	Number	Per cent
Manufacturing	21,663.00	6.19
Farming	242731.00	69.41
Service	78,507.00	22.45
Mining	6,799.00	1.94
Total	349700.00	100.00

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 ETL.

References

Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103(484):1481–1495.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008(10):P10008.

Csardi, Gabor and Tamas Nepusz. 2006. "The igraph Software Package for Complex Network Research." *InterJournal* Complex Systems.

Englebert, Pierre. 2005. "Compliance and Defiance to National Integration in Barotseland and Casamance." *Africa Spectrum* 40(1):29–59.

Erdmann, Gero. 2007. "Ethnicity, Voter Alignment and Political Party Affiliation - An African Case: Zambia." *German Institute of Global and Area Studies Working Papers* 45.

Fortunato, S and M Barthelemy. 2007. "Resolution Limit in Community Detection." *Proceedings of the National Academy of Sciences* 104(1):36–41.

Gibson, Clark C and Barak D Hoffman. 2013. "Coalitions not Conflicts: Ethnicity, Political Institutions, and Expenditure in Africa." *Comparative Politics* 45(3):273–290.

Haanyika, Charles M. 2008. "Rural Electrification in Zambia: A Policy and Institutional Analysis." *Energy Policy* 36(3):1044–1058.

Haythornthwaite, Caroline. 2005. "Social Networks and Internet Connectivity Effects." *Information, Communication & Society* 8(2):125–147.

Knack, Stephen, F Halsey Rogers and Nicholas Eubank. 2011. "Aid Quality and Donor Rankings." *World Development* 39(11):1907–1917.

Kramon, Eric and Daniel N Posner. 2013. "Who Benefits from Distributive Politics? How the Outcome One Studies Affects the Answer One Gets." *Perspectives on Politics* 11(02):461–474.

Lancichinetti, Andrea and Santo Fortunato. 2009. "Community Detection Algorithms: A Comparative Analysis." *Physical Review E* 80(5):056117.

Miritello, Giovanna, Esteban Moro, Rubén Lara, Rocío Martínez-López, John Belchamber, Sam G B Roberts and Robin I M Dunbar. 2013. "Time as a limited resource: Communication strategy in mobile phone networks." *Social Networks* 35(1):89–95.

Moody, James and Douglas R White. 2003. "Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups." *American Sociological Review* 68(1):103–127.

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*.

Onnela, Jukka-Pekka, Jari Saramäki, Jorkki Hyvönen, Gabor Szabó, David Lazer, Kimmo Kaski, János Kertész and Albert-László Barabási. 2007. "Structure and Tie Strengths in Mobile Communication Networks." *Proceedings of the National Academy of Sciences* 104(18):7332–7336.

Osei-Hwedie, Bertha. 1998. "The Role of Ethnicity in Multi-Party Politics in Malawi and Zambia." *Journal of Contemporary African Studies* 16(2):227.

Phiri, B J. 2006. *A Political History of Zambia: From Colonial Rule to the Third Republic, 1890-2001*. Trenton, New Jersey: African World Press, Inc.

Posner, Daniel N. 2004. "Measuring Ethnic Fractionalization in Africa." *American Journal of Political Science* 48(4):849–863.

Posner, Daniel N. 2005. *Institutions and Ethnic Politics in Africa*. Cambridge: Cambridge University Press.

Rabe, Marlize. 2016. "Family policy for all South African families." *International Social Work* 60(5):1189–1200.

Rojo, Guadalupe and Erik Wibbels. 2014. "Clientelism, Public Services and Elections in the Slums of Jaipur and Udaipur." *Presented at Stanford Comparative Politics Seminar* .

Roodman, David. 2005. "An Index of Donor Performance." *Center for Global Development Working Paper Series* .

Sardanis, Andrew. 2014. *Zambia: The First 50 Years*. London: I.B. Tauris.

Taylor, Scott D. 2006. *Culture and Customs of Zambia*. Westport, Connecticut: Greenwood Press.

Traag, V A, P Van Dooren and Y Nesterov. 2011. "Narrow Scope for Resolution-limit-free Community Detection." *Physical Review E* 84(1):016114.

USAID. 2008. Zambia: Water and Sanitation Profile. Technical report.

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.

Yen, Luh, Denis J Vanvyve, Fabien Wouters, François Fouss, Michel Verleysen and Marco Saerens. 2005. Clustering Using a Random Walk-based Distance Measure. In *European Symposium on Artificial Neural Networks*.

Zambia 2010 Census of Population and Housing. 2012. Technical report.