Digitally networked grassroots

Social media and the development of the movement for black lives and immigrant rights movement in the United States

van Haperen, S.P.F.

Publication date
2019

Document Version
Other version

License
Other

Chapter 11

The Nuts and Bolts of Computational Social Science: Methodological Reflections on Fitting Power Law Distributions[190]

Sander van Haperen

190 A version of this chapter is under review for publication at the time of writing, August 26, 2019.

Abstract

The combination of computational methods and data from social media allows for powerful ways to study the development of social movements. The implementation of computational approaches, however, presents us with epistemological and technical challenges. For instance, we can use these methods to describe empirically how prominence develops in networks of millions of users on Twitter (see Chapter 10). Such empirical descriptions are a first analytical level. A second analytical level concerns epistemology: how we use these computational methods to arrive at empirical findings.

I will argue in this chapter that seemingly inconspicuous details involved in the implementation of computational methods limit the validity of findings. In other words: to unpack the black boxes of social movement development, we have adopted methods that are in themselves black boxes (cf. Goldstone, 2015; Törnberg and Törnberg, 2018; Tufekci, 2014). This is brought to light by detailing the limits of a seemingly straightforward measurement. Power law modeling is a particularly interesting case, as a hallmark of network analysis and the subject of ongoing debates (Broido and Clauset, 2019). *Practically, it is easy to make power law distributions fit empirical data by using a software package like igraph in R* (Csardi and Nepusz, 2006). *Under the hood, however, the details of this measurement can be quite overwhelming, as I hope to demonstrate. It is not straightforward to determine how inputs are transformed into outputs. The convenience offered by this computational instrument obscures nontrivial technical decisions and underlying assumptions that potentially have a large impact on the findings.*

To address these limitations of computational methods, I discuss a relational approach that accounts for the experience of social life in relation to network structures (Crossley, 2015; Diani and McAdam, 2003; Elias and Scotson, 1994; Emirbayer, 1997; Uitermark, 2010). *Its basic tenet is that structural relations and community practices are interdependent. Practically, the relational approach proposed here integrates qualitative inquiry in specific contexts with the selection and refinement of network measurements.*

Introduction: "You don't understand Twitter, my friend"

Setting out to examine the development of digitally networked movements led me to collect data from Twitter. Adopting the tools of network analysis, I treated mentions and retweets as indicative of social relations. Conducting fieldwork for this study allowed me to interview respondents who had emerged as significant from among the millions of nodes in the networks I concurrently examined using computational methods. Often, while trying to make sense of the development of social movements, I discovered insights that went against my understandings of digital networking based on network measurements. Every opportunity they were kind enough to allow me, my respondents helped further my inquiry by challenging my assumptions.

To one respondent I suggested that she occupied a central position in the network of activists in Los Angeles, based on the mentions she received. Not impressed, she told me, "You don't understand Twitter, my friend." In part, I believe she was being humble, downplaying her importance and position of prominence, for which she demonstrably was working hard. Explaining that she had once been important in "certain parts of the old LA [Los Angeles] left," she had deliberately distanced herself from that scene. Disillusioned by quarrelsome old comrades, she "moved on to social media" to engage in ways that were more about "exchanging ideas." Now, she "builds followers like everyone else," something at which she was quite adept, her centrality in my network analysis suggested. What had been abstract nodes on a graph were now speaking to me, refuting the relational positions I had attributed to them. How to be certain that the network topology was not the result of noisy data or a methodological artifact? In other words, I faced a dissonance between empirical findings and epistemological approach that needed to be reconciled. It made me question in two ways how to make sense of network analysis.

First, computational methods pose both obvious and inconspicuous technical challenges. Some of the tools I used were black boxes. It is easy to produce results with a well-established software package like 'igraph' for R (Csardi and Nepusz, 2006). With a minimum of coding or statistical skills, it is easy to produce interpretable results with pre-programmed packages. For example, the widely employed igraph offers many convenient functions, excellent documentation and skilled support communities. However, it is far from straightforward to determine how specific functions in igraph transform input into output. In practice, every

line of code is a research decision that invites examination. As a result, the process of producing interpretable findings involves many decisions at every step that are far from trivial, from collecting to processing to analyzing to presenting data. Inconspicuous function parameters, typing errors, or small oversights that do not necessarily interrupt runtime can have large implications for findings, a problem further exacerbated by large data sets. One example is the way in which a certain function in the popular igraph package is implemented, demanding specific conditions for how it returns particular *p* values of a statistical test. This determines the significance (literally) of the output from that function, but without examining the source code, there is no way of knowing this. Even though the most skilled programmers often err (Raymond, 1999), current research practices make reflection on such intricacies rather rare in computational social science (Broido and Clauset, 2018; Chan et al., 2019; Poursabzi-Sangdeh, Goldstein, Hofman, Vaughan, and Wallach, 2018; Sharma, Hofman, and Watts, 2019).

Secondly, while computational methods and digital data allow us to examine networks in unprecedented ways, interpretation of the measures they produce is not straightforward. These epistemological, second-level challenges potentially lead empirical analysis astray. For example, receiving many mentions can be conceptualized as indicative of prominence, and a leader's prominence may be measured by examining centrality in the networks of mentions made. From the vantage point of computational methods, a node does not need to be aware of taking up a central relational position while fulfilling such a network mechanism. In the Los Angeles network, the abovementioned respondent connected a group of activists primarily interested in environmental issues with a group of Bernie Sanders supporters, by retweeting content from both groups. The structural position of this node in a network is analytically interesting in its own right. It tells us that the respondent plays a role in sharing information between different, otherwise unconnected, groups of people, facilitating the diffusion of information, even if she refuted that doing so was important. It is difficult to reconcile analytically individual perspectives with the scope of structural network analysis, particularly when considering the millions of people connected to a social movement. Arriving at the abstractions typical of computational methods (for example, inequality in networks) necessitates bracketing, rather than contextualization, of subjectivities. While difficult to resolve analytically, contextualization further enhances the power of computational methods by helping us to ask better questions and guiding interpretations of network measures.

Power law distributions specifically offer an interesting case study that features in several chapters of this dissertation, of significance because of the ongoing debates about the measurement of this hallmark of network analysis (Barabasi and Albert, 1999; Broido and Clauset, 2019; Pastor-Satorras and Vespignani, 2001; Rainie and Wellman, 2012; Watts and Strogatz, 1998). The debate powerfully illustrates an incremental development of knowledge from critical interrogation of long-established, seminal findings (Broido and Clauset, 2019). In my research, the measure is used to answer a seemingly straightforward question: how concentrated is leadership in the digitally networked movement #blacklivesmatter? Drawing on digital data and statistical methods, Chapter 10 set out to answer this question using power laws as a measure of concentration. That chapter may be read as a case of a straightforward research inquiry which yielded concrete results that inform an argument and make a theoretical contribution to the literature. In this chapter, those results are used as a case study. As will become clear, the process of *conducting* that inquiry was not at all straightforward. While focusing here on power laws, recent research demonstrates similar pitfalls hold for community detection (Traag, Waltman, and van Eck, 2018), and the same is likely the case for topic modeling, sentiment analysis, and other established computational tools (Chan et al., 2019).

In short, while computational methods offer powerful new tools to analyze the development of social movements, the implementation of these tools poses technical and epistemological challenges. This chapter contributes a critical reflection on the limitations and possibilities of the computational relational approach developed for this dissertation. It asks: How can we use power laws to study and make sense of "networks of outrage and hope" (Castells, 2012)? As this chapter demonstrates, it turns out that a convenient and seemingly straightforward network measure involves many nontrivial decisions with significant impact on the interpretation of findings. A relational approach is proposed to address these limitations.

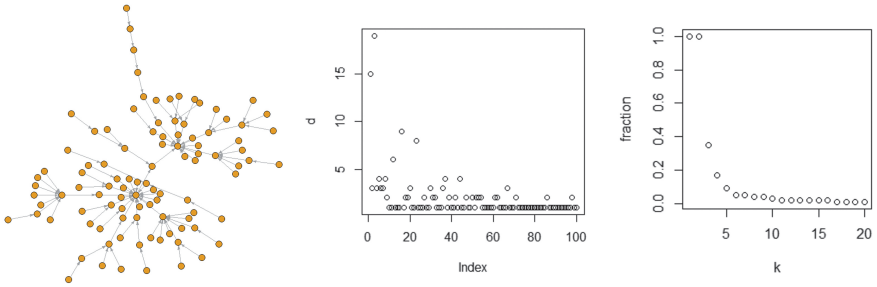
Power Laws: Assumptions and Implementation

11

Fully understanding the theoretical background of statistical measurements, such as power law distributions, is a challenge. Nevertheless, power laws have become a mainstay of network analysis (Barabási, 2015; Clauset et al., 2009). As the central

tenet of scale-free networks based on the theory of phase transitions, they are key to theoretical notions of diffusion, contagion, brokerage, and tipping points. One must, however, pose the question: Is a power law exponent the right metric to measure the concentration of prominence in a social movement network? Imagine a network of 100 people, who have a certain number of ties among them. Some nodes have a single tie, whereas others have two or more ties. This is a distribution, which, for a generated semi-random network, could look like this:

Figure 11.1: Distributions in a simulated network



From left to right: graph, degrees, and cumulative degree distribution[191].

While most nodes in this network have a single tie to another node, there is one node that stands out due to having 20 ties. For the network graph shown on the left side of Figure 11.1, the central picture depicts the researcher's observations: the number of ties for each 100 observations. If we want to know how unevenly ties are distributed among these nodes, on the right side of Figure 11.1 is the distribution of cumulative degrees: the fraction of nodes (from 0 to 1) with a degree smaller than *k*. Here, we see that the node that stands out with its 20 ties represents a small fraction of the network. Statistically, this can be understood as the probability of selecting a node at random from a network with a degree lower than *k*. Because only one of the 100 nodes has a degree of 19, the probability of randomly picking a node with a degree lower than 19 is 99%. In this example, the nodes turn out to have one of only nine different degrees: 1, 2, 3, 4, 6, 8, 9, 15, and 19. As it turns

191 Generated with igraph in R with a Barabasi-Albert game, a simple stochastic algorithm that follows: P[i] ~ k[i]^alpha + a. Note: This algorithm by definition produces a degree distribution fitting a power law exponent.

out, 65 of the 100 nodes have only a single tie, 18 nodes have 2 ties. Thus, there are no nodes with 5 ties, no nodes with more than 19 ties, et cetera. To know exactly how uneven this distribution is, it makes sense to fit this distribution to a power law model.

In its most basic form, a power law is a function of two variables, in which a change in one variable produces a proportional change in the other variable. In the case of a probability distribution with the scaling parameter *a*, this takes the form of:

$$p(x) \propto x^{-a}$$

Applied to the distributions of ties in a network, the derivative of this function is an exponent that can be interpreted as a measure of unevenness. In our case, as the node degree increases, the probability of randomly selecting a node with a high degree decreases. A key property of power law distribution is scale invariance: the exponent applies to any point in the relationship between the variables, so that the scaling of one proportionally scales the other; in other words: it is linear on a log-log plot. Referring back to Chapter 9, what we want to learn is how uneven the distribution of ties is in relation to the number of people whose tweets are plotted. In short, for our purposes, the properties of power laws help to better understand inequalities in digitally networked movements.

There are two ways to model a power law distribution, with a different formula for continuous data and one for discrete data. Conceptually, the question of whether network edges represent continuous or discrete data is nontrivial. The convention in many network studies is to treat non-weighted edges as discrete and weighted edges as continuous. One could argue that all observable data are, in the end, discrete (Pitman in Aitkin, 1979, p. 1). Nevertheless, an epistemological problem remains, namely, that it is not certain what the data represent. This problem is compounded when one calculates the exponent as a probability fraction in the form of a cumulative degree distribution, producing continuous data. For an illustration of the advanced methods currently being developed in other fields where researchers are facing this conundrum, see, for example, works in epidemiology,[192] the modeling of cellular processes, (Jamshidi, 2012) and deep learning and neural networks (Klambauer, Unterthiner, Mayr, and Hochreiter,

192 https://nbviewer.jupyter.org/github/simoninireland/cncp/blob/master/epidemic-network.ipynb, accessed August 5, 2019.

2017). For discrete data, which our network application arguably examines, the exponent can be approximated with maximum likelihood estimators:

$$\frac{\zeta'(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i$$

This takes the ζ zetas over *n* as the number of observations, with *xi* as values higher than the threshold *x*min. The logarithmic derivative produces an estimate of the logarithms of the sum over the threshold, normalized to the number of observations. This equation may be solved by maximizing its likelihood function:

$$L(\alpha) = -n \ln \zeta(\alpha, x_{\min}) - \alpha \sum_{i=1}^{n} \ln x_i$$

This formula reduces to the Riemann zeta function. Formulated in 1859, the Riemann hypothesis is widely recognized as the most important and difficult unsolved questions in mathematics. Practically, as discussed below, this is resolved by approximating convergence, rather than producing an exact numerical calculation. Having established an estimate of the exponent, we need to determine standard error in order to evaluate the remaining uncertainty. In other words: what is the probability of the observed dataset, given a model based upon an estimated exponent? To assess this probability, we can calculate a confidence interval with the quadratic of a maximum log-likelihood:

$$\sigma = \frac{1}{\sqrt{n\left[\frac{\zeta''(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})} - \left(\frac{\zeta'(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})}\right)^2\right]}}$$

Here, we rely on the Gaussian behavior of maximum likelihood estimation, in accordance with the central limit theorem (as shown in Gilbert, 2000[193]). The result is the log-likelihood. From all the distributions that could have been obtained from the estimated exponent, how likely is it that this model produced the data observed? In most practical cases, a good fit of the power law to the distribution does not apply to the entire distribution, so it is commonplace to exclude part of

193 https://projecteuclid.org/euclid.aos/1016120368, accessed August 5, 2019.

the tail to make the distribution fit[194]. To that end, a threshold that excludes values below a certain value may be used.[195] If it is unclear what the threshold (*xmin*) is, it must be estimated. Weighing different approaches, Clauset et al. (2009) propose quantifying the distance between the empirical observations and the modeled power law distributions. By varying the threshold incrementally, the optimal *xmin* is determined. To do so, a Kolmogorov-Smirnov test can be used, reweighted to measure distance across the 0-1 range uniformly:

$$D^* = \max_{x \geq \hat{x}_{\min}} \frac{|S(x) - P(x)|}{\sqrt{P(x)(1 - P(x))}}$$

The threshold that minimizes distance (D) serves as the estimated threshold, xmin. In order to assess the uncertainty of this estimation, we can compare it to a large number of synthesized iterations drawn from the observed data. Bootstrapping by the standard deviation of those randomizations yields a comparison that quantifies the level of agreement with the lower threshold in the observed data.

Following the reasoning outlined thus far, it is possible to fit a power law model to *any* dataset. We have not yet assessed how well the model fits the observed data. Having estimated the parameters of the power law model, we next need to assess how plausible it is that the model represents our empirical data. Clauset et al. (2009) propose bootstrapping to test the hypothesis that the empirical data actually fits a power law model and, to that end, employ the Kolmogorov-Smirnov test mentioned above. A *p* value helps in assessing whether the hypothesis of the power law can be confirmed when compared to other models.[196]

The Practical Implementation of Power Law Measurements

Practically, there are different ways to calculate a power law exponent, and programming the code for this calculation is nontrivial. Rather than coding from scratch, many researchers use code from packages or 'libraries.' The most

194 Note that rather than "good fit", used as short-hand here, a more accurate description would be: ""it is very unlikely that the data could have been generated from the hypothesized distribution" (in our case, a power-law). A high p-value _roughly_ means that 'the data may have come from the hypothesized distribution'; however, there could be alternative distributions that can describe the data just as well." (Nepusz, correspondence, August 2019).

195 Incidentally, this means that a good power law fit typically applies only to nodes with more than a few ties, excluding substantial parts of a network.

196 Note that the formulation of hypotheses takes a specific form here, the practical implications of which will be discussed below.

widely used package to calculate power laws for R is called igraph.[197] This package "provides handy tools for researchers in network science" (Csardi and Nepusz, 2006, p. 1). It includes the function "fit_power_law," which "fits a power-law distribution to a vector containing samples from a distribution" and was written by Tamas Nepusz and Gabor Csardi. As discussed above, estimation of the power law exponent is based on different theoretical assumptions for continuous vis-à-vis for discrete data. Accordingly, practical implementation in igraph employs two distinct procedures. Programmatically, the distinction is necessarily one of integers and floats.[198] One pragmatic way to decide whether to use the discrete or continuous procedure would be to argue that, seeing that our operationalization of network relationships does not suggest that fractions of edges are meaningful, it is reasonable to treat the data as discrete. The fit_power_law function in the R igraph package defaults to continuous for vectors containing at least one non-integer, if no deliberate choice is made.

Practically, however, the function deals with data input in specific ways. For example, while necessary for effectively estimating lower bounds, it is not required to sort the vector, since the package implementation sorts automatically. Likewise, it offers convenient options, such as forcing discrete data into a continuous implementation, either by truncating them or through the function parameter "force.continuous." It is tempting to do this because the equation is easier (and faster).[199] However, this is problematic because it misrepresents data and produces a poor model fit, particularly for small data sets[200] (n =< 50) and a threshold less than 6. This is not noted in the package's help file; nor is it obvious from the inconspicuous output of the function. In the case of networks, we assume people have one or more ties, without fractions. For ties in a Twitter network, this lower limit is problematic as a large share of nodes may be expected to receive fewer than six mentions.

---

197 Colin Gillespie's poweRlaw package provides an excellent alternative. Although less frequently used than igraph, it is dedicated, well-documented, and well-maintained (https://cran.r-project.org/web/packages/poweRlaw/vignettes/b_powerlaw_examples.pdf, retrieved July 29, 2019).

198 Incidentally, this is a nontrivial matter in computing architecture and the key source of the infamous history of competition between Apple and Microsoft. Wozniak did not think it necessary to add floating point math to the Apple operating system, so as to save time, and he thus opted for performance over precision (https://gizmodo.com/how-steve-wozniak-wrote-basic-for-the-original-apple-fr-1570573636, retrieved July 30, 2019). As a result, Microsoft's BASIC needed to be licensed later.

199 This could be implemented in R with something like "n/sum(log(round(z)/xmin))."

200 According to the plfit.r implementation it is (n <100) (Dubroca, 2008), and (n =< 50, according to Clauset et al. (2009).

A related point of confusion about the transformation of empirical data in the function arises from the still-widespread use of least-square estimation on the basis of cumulative degree distributions. The igraph implementation uses maximum least squares to estimate the scaling parameter for the observed values, rather than the logarithm (as is common in least-square approaches). Thus, it is possible to easily estimate the cumulative degree distributions, as is often done, presumably as a misconception of the underlying equation. The igraph function "degree.distribution"[201] simply computes a histogram of the observed degree counts (density) in a graph object, the cumulative parameter in this case serving to normalize values to fractions of 1.[202] If one assumes there is a linear scalar when fitting a power law, the exponent will be different for fitting cumulative from that for fitting non-normalized values. A cumulative degree distribution naturally follows a power law, so that outcomes of the Kolmogorov-Smirnov statistic will generate *p* values close to 1 for any cumulative degree distribution. Applying the fit_power_law function to our example of 100 nodes with 198 ties, the degree distribution versus the cumulative distribution (see figure 11.1) return exponents of 2.58 and 1.64 respectively. Again, the output is inconspicuous. Both have significant *p* values for the Kolmogorov-Smirnov test, so that the power law can be accepted as a similarly good fit,[203] while the cutoff is estimated differently and returns different log-likelihood estimators (-58.68 and 2.92 respectively). Practically, this implies that the exponent can produce a difference of a factor of 10, despite the. fact that the distribution is empirically identical. When empirically examining social movements, this severely limits the generalizability of power law distributions, particularly across studies.

These log-likelihood estimators are calculated in the function by using a maximum likelihood estimation to make a power law fit a degree distribution. Put simply, this estimates the curve of the distribution of the number of ties for all nodes, so that we can compare this to other distributions. Maximization of the likelihood function allows us to estimate this alpha value. In the igraph package, this is implemented through a C call that initially calculates the exponent simply

201 https://rdrr.io/cran/igraph/src/R/structural.properties.R, retrieved May 28, 2019.

202 Interestingly, the default for the function hist() to determine how to break values takes the max value divided by the number of observations. The degree.distribution function overrides this default by forcing breaks from -1 to the maximum value, so that any observations of 0 or 1 are counted, not binned with 2. While introducing a 0 to the vector, the effect is nil as the power law equation will later take the log of 0 and 1, which is 0.

203 i.e., we can not rule out that the data comes from a power-law distribution, but it may still come from alternative distributions.

as the cumulative sum of the natural logarithm of the values.[204] This exponent is then processed using the Broyden-Fletcher-Goldfarb-Shanno optimization function of the lbfgs library (Nocedal and Okazaki, 1990), a quasi-Newton code that approximates the best fitting scalar in an inverse hessian matrix over a number of iterations (not the bfgs function as the R igraph documentation suggests[205]). To do so, the lbfgs library relies on the gsl_sf_hzeta function of the zeta.h library, an implementation of the generalized Hurwitz zeta function (GNU Scientific Library, 2007), calculated as:

$$\text{zeta(s,q)} = \text{Sum[(k+q)^(-s), \{k,0,Infinity\}]}$$

The lbgfs method is based on the assumption that the objective function f is continuously differentiable twice (Liu and Nocedal, 1989, p. 21). In practice, that assumption results in imposing a power law distribution to part of the data in order to estimate an exponent. For networks, we might assume an exponent between 1 and 10, as empirical networks typically obey a power law between $2 < x < 3$. Instead, in the C implementation (on which igraph relies) an infinite upper limit of 1e10 is simulated, and a value close to the observed distribution approximated. While programmatically convenient, the unrealistic assumption on which estimation is based in the function means that any data can be fit to the model to generate outcomes.

However, an estimated power law exponent is not a good fit for the entire distribution of most empirical data, so it is commonplace to exclude part of the tail to fit the distribution. This means that specific parts of the network are disregarded in the distribution model. Nodes with fewer than a specific number of ties are excluded. If it is unclear what the threshold (*xmin*) is, it needs to be estimated. The fit_power_law function in the igraph package offers two algorithms to perform this: plfit and MLE. By default, the program uses the newer implementation, plfit.[206] While both adopt maximum likelihood estimation as a current best practice (rather

204 I do not discuss continuous data here because my data concerns (discrete) network ties. The continuous function is more straightforward and normalizes values to n and the threshold xmin (Nepusz, 2010).

205 The reason for this is likely pragmatic: the lbfgs implementation uses less memory than bfgs (Liu and Nocedal, 1989).

206 While both make use of maximum likelihood estimation, the older implementation relies on the mle function of the stat4 package. It is not clear why this function was deprecated. The "The plfit library is an efficient implementation of the method published by Clauset, Shalizi and Newman" (Nepusz, correspondence, August 2019).

than a least-squared linear regression), plfit allows for the manual thresholding of the minimum degree at which to apply the power law.[207] The probability of the observed dataset generated using a power law model with the estimated exponent is assessed by calculating a confidence interval with the quadratic of the maximum log-likelihood. In the C implementation of plfit, this is calculated as:

```
result = - alpha * result - m * log(gsl_sf_hzeta(alpha, xmin));
```

where alpha is the estimate of the exponent yielded by the optimization function. We can now compare the log-likelihood ratio between the fitted power law and alternative exponential models. The C implementation of this log-likelihood ratio in the igraph package employs the Hurwitz Zeta function.[208] As a result, interpretation of *p* values output is less straightforward than might be expected. Clauset et al (2009) interpret high *p* values resulting from the Kolmogorov-Smirnov test as confirmation of the hypothesis, which may be counterintuitive for many sociologists used to ruling out null hypotheses, typical in regression analysis. The formulation of the hypothesis prescribes that a *high p* value suggests that a power-law distribution cannot be ruled out. "Normally one then considers low values of *p* to be good, since they indicate that the null hypothesis is unlikely to be correct. Here, by contrast, we use the *p* value as a measure of the hypothesis we are trying to verify, and hence high values, not low, are 'good.'" (Clauset et al., 2009, p. 678, fn #8). This is implied but may not be obvious from how *p* values are reported in igraph, defining the *p* value in the help file as,

> Numeric scalar, the *p* value of the Kolmogorov-Smirnov test. Small p values (less than 0.05) indicate that the test rejected the hypothesis that the original data could have been drawn from the fitted power-law distribution (Nepusz and Csardi, 2003 n.p.)

This is further complicated by the fact that Clauset et al. (2009) suggest calculating two distinct *p* values: one for the Kolmogorov-Smirnov statistic and one for likelihood ratio. For the latter, a small *p* value is "good," as it indicates that a direct comparison of models is trustworthy. Furthermore, note that Clauset et al. suggest

207 If not provided, the plfit library uses the Kolmogorov-Smirnov statistic to estimate lower bounds, in accordance to the method proposed by Clauset et al. (2009). This statistic is discussed below.

208 Defined as: zeta(s,q) = sum_0^infty (k+q)^{-s}. See discussion above.

$p < 0.1$ because more leniency would "let through some candidate distributions that have only a very small chance of really following a power law" (2009: 17). At this confidence interval, high p values are likely to result spuriously from small-n data. For data with $n = < 100$, it is not possible to accurately determine p values, so that the hypothesis of a power law cannot be confirmed.[209] Moreover, because of how zetas are calculated in the plfit package, only observations above the threshold xmin are considered. This means that, in order to arrive at a statistically significant and robust estimation of the power law exponent, at least 100 observations need to be above the given threshold. In other words, in the case of our conjectured network, assessment of the p-statistic is problematic when there are fewer than 100 observations at a degree higher than 6. The opposite is also true, and perhaps more worryingly: nominally robust outputs are easily produced for empirical networks with at least 100 nodes, while those outputs may exclude, say, 50,000 nodes from the fitted model. Thus, in order to correctly evaluate validity and plausibility of any power law exponent, we thus need to report both the associated Kolmogorov-Smirnov p value, as well as n and xmin.

Implications for the Current Study: The Transformation, Presentation, and Interpretation of Data

Is consideration of all the arcane technical details laid out in this chapter really necessary, if we are interested in social movements? The purpose of presenting the details of power law models was to demonstrate that it is not straightforward to determine how inputs are transformed into outputs in the igraph package. This is problematic because seemingly inconspicuous decisions can have a significant impact on findings: the devil is in the details.

A First Illustration of Nontrivial, Inconspicuous Details: Data Collection and Transformation

Not only do the intricacies of the procedure for modeling power laws potentially impact findings, but of course the findings crucially depend on data collection itself. The discussion thus far has focused on the power law and using igraph, as one of many technicalities that shape findings. A brief discussion of data collection

209 Note that accurate estimation of the xmin value requires "about 1000 or more observations" (Clauset et al, 2009: 672) in the n-tail part of the distribution.

serves to further illustrate how seemingly trivial decisions have significant impacts on findings.

There are different ways to collect data from Twitter. This dataset was constructed by monitoring the Twitter API platform (Twitter Inc., 2014). Of the various endpoints Twitter API offers, tweets were selected according to their hashtags, rather than specific users. Based on preliminary exploration, variations on the hashtag #blacklivesmatter were used to query the API, such as "blacklivematters" and "blacklivematter." Elements from these data were conceptualized as ties in a Twitter network: we inferred a relation between accounts when one user mentioned another. The accounts represented different kinds of actors, such as individuals, media, and organizations. Many accounts show signs of automation, suggesting non-human actors. To prepare the data for network analysis, it was transformed in various ways. For instance, I used SQL in combination with regular expressions for identifying and selecting mentions from the raw texts of tweets.[210] In addition, a range of operations were involved in identifying supporters and opponents on the basis of tweets, discussed in detail in the methodology section of Chapter 6. In short, using content analysis, I developed a semi-supervised method involving the interpretation of progressive samples from the data, which then informed my large-scale algorithmic classification. This in itself involved a range of intricacies. How was the sample constructed? Are unreciprocated ties included? Is the measure calculated for indegree or outdegree or both, for *targets* or sources, with what minimum support score? And so on. It is clear that each of these questions warrants closer inspection. I will not do so here; the point is that such intricacies are so easily obscured.

Collecting data from the streaming API introduced biases outside the control of the researcher because it returned a sample, rather all the tweets that include the hashtag. The free, streaming API returns about 1% of tweets with a particular hashtag, in comparison to the commercial, "firehose" service Twitter offers that supposedly returns complete data (Twitter Inc., 2014). While it is not clear exactly how the data are sampled, it is possible to assess data biases by comparing

210 Regular expressions are a powerful tool, but infamously difficult to handle. "Some people, when confronted with a problem, think 'I know, I'll use regular expressions.' Now they have two problems." (Jamie Zawinsky as cited in Friedl, 2006). Crucially, the problem derives from linguistics and the use of symbols in computer science, which is beyond the scope of this chapter (Chomsky, 1959). A key concern are edge cases in complex n-grams. Although primarily interested in simple, well-defined unigrams ("@"), I relied on trial and error in combination with manual examination and data validation.

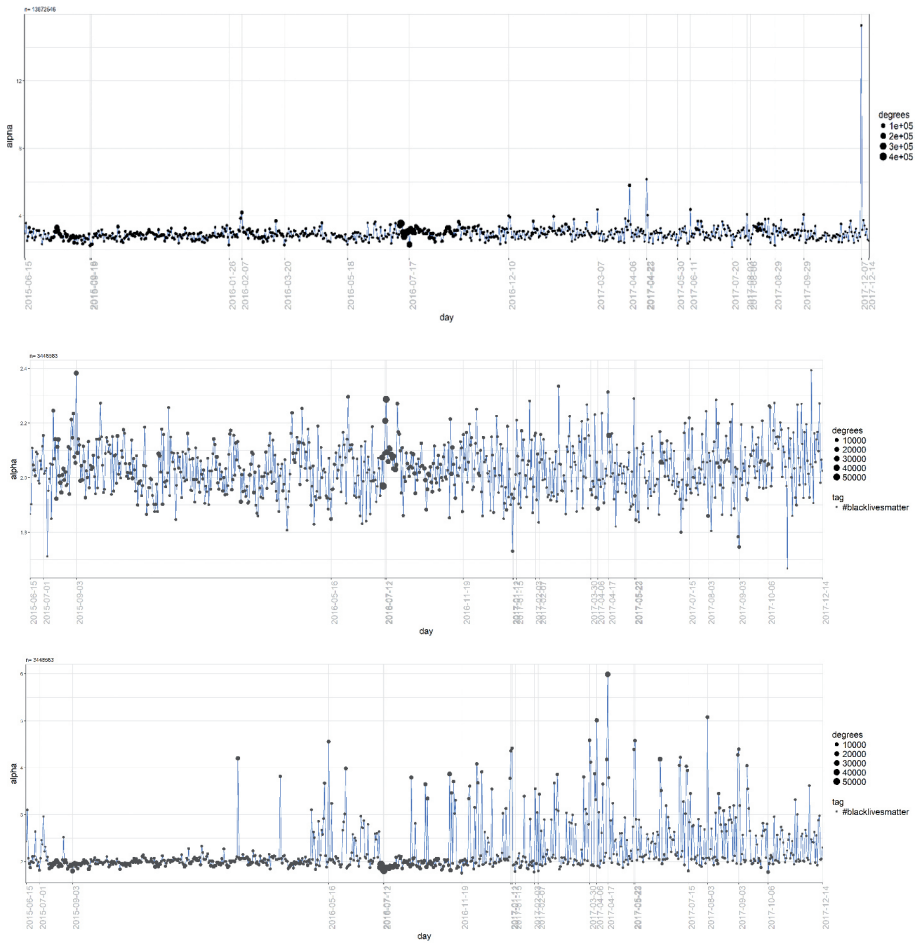
various collection strategies (González-Bailón, Wang, Rivero, et al., 2014; Tufekci, 2014a). Estimations of network centrality measures in the data returned by the streaming API closely matched those for firehose data in sufficiently large data sets (Morstatter et al., 2013). The sampling biases were addressed by aggregating data, normalizing over an extended period over time. This shifted emphasis to relative changes as a unit of observation, instead of as the absolute exponent.

Aggregating data introduced a new concern: There was no natural temporal break at which to aggregate tweets. On the one hand, aggregating data by month shows a more pronounced trend towards a concentration of prominence later in the period under observation. The outliers were evened out so that the range of the alpha was smaller. Thus, daily volatility was no longer visible, emphasizing instead a "less leaderful" movement from June 2016 onwards. It should be noted that an exponent range between 1.70 and 1.90 can be considered very evenly distributed, when compared to other networks which typically range between 2 and 3. Obviously, selecting the period to observe crops the picture and how we interpret it. Data collected and depicted only from the later period, between late 2016 and early 2017, suggested (slightly) more concentration of prominence, but also a trend towards less concentration. By contrast, the period between September and December 2015 suggested more stability than did the overall picture. On the other hand, qualitative inquiry suggested that discussions were strongly tied to the daily news cycle, suggesting 24 hours as a suitable period into which to aggregate data. Some respondents argued that "Twitter-time" spans shorter periods: they saw a few hour delay in a discussion on Twitter as a sign that someone is not "in the loop," not on top of things. Thus, while aggregating days may be acceptable, the point is that there is no perfect unit of measurement: depending on how the data is aggregated, modeling power law distributions will produce different findings.

A Second Illustration of Nontrivial, Inconspicuous Details: Refinement and Presentation

Second, the measure itself and the presentation of findings were refined in conversation with the data. To a large degree, this affected the conclusions: Different representations of the data in illustrates some of the less obvious pitfalls discussed in the previous sections (Figure 11.2).

Figure 11.2. Three representations of the same power law model



The power law exponent, by day, for: all nodes in the network (top); only among supporters (middle), and non-cumulatively among only supporters (bottom).

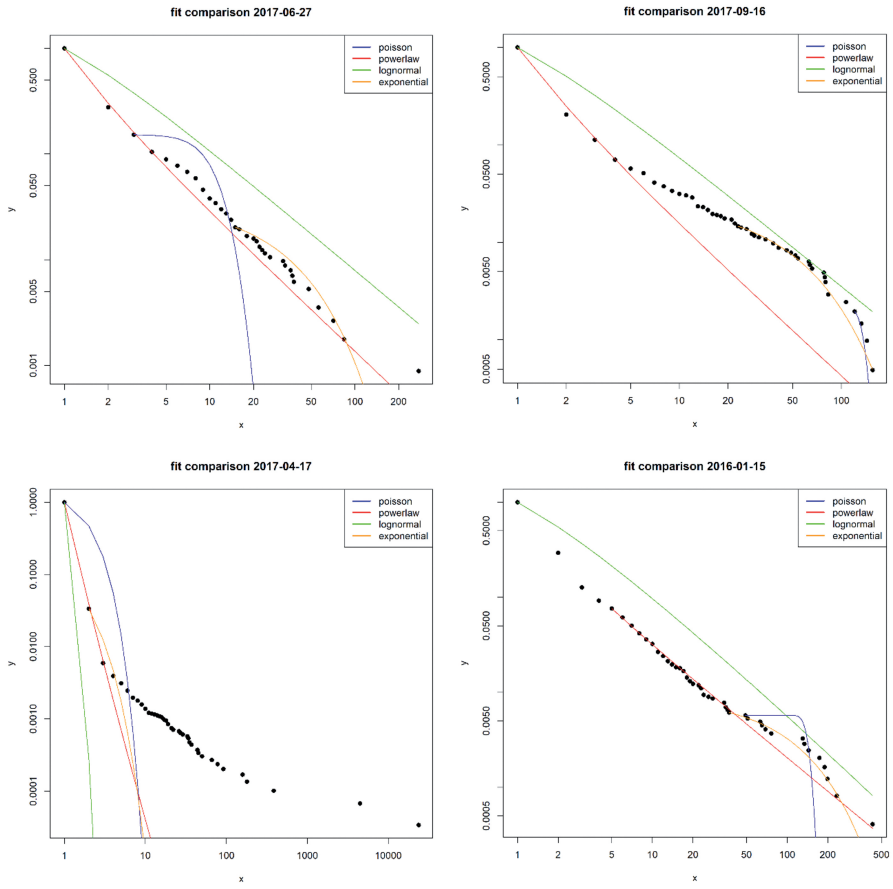
The graphs illustrate different ways to present the same data. If the first impression is one of a stable exponent, this is because a single outlier changes the scale drastically. While an obvious outlier when presented as one, examining data on a single day or according to a different timeframe obscures this fact. The outlier on December 7, 2017 was caused by @realdonaldtrump; obviously not someone we think of as a leader of Black Lives Matter, but this can only be determined by contextualization. As stated above, filtering out opponents became important

early on and the bottom graph depicts only supporters. There are still outliers (note that the x-axis is a logarithmic scale), but the pattern and daily fluctuations are much more pronounced. The bottom graph depicts the same data, but with exponents calculated non-cumulatively, a practical technicality that can be easily overlooked, as discussed before in the section on using igraph. Note that the same differences hold for opponents; it is not the case that supporters and opponents together provide a more stable pattern. Inconspicuous technicalities and different visual representations of the same data obscure key findings.

Checking for robustness allowed for further refinement of the measurement. Figure 11.2, above, was based on a naïve calculation of the power law exponent for daily slices of the network. It reports, regardless of sample size, lower bounds, or *p* values, the power law exponent among supporters. However, not all the alphas depicted are representative or significant of a power law distribution (Clauset et al., 2009). Each point is calculated based on the degrees generated during a single day. Different scenarios can account for the observed distributions, examples of which are depicted in Figure 11.3.

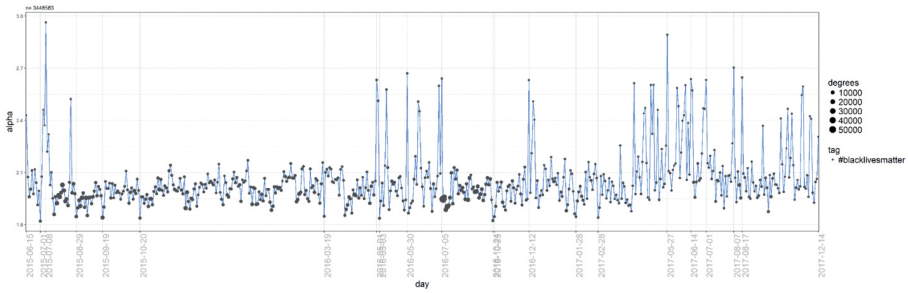
One of the outliers depicted in Figure 11.3 is the alpha 5.99 for April 17, 2017. This distribution is depicted on the bottom left of Figure 11.3. As one might suspect, the low *p* value (0.0294) suggests the power law model is a poor fit. None of the other four considered models proved a good fit either. Thus, while the distribution suggests there is some degree of unevenly distributed prominence, a direct comparison with the alphas from other days would be naïve. To correct for this, the *p* values need to be considered. Another scenario, according to which the power law is a good fit, but only with a high lower threshold, is depicted on the bottom right. This means that the large majority of the network is completely disregarded. In such cases, the alpha is not a valid measure of claims about the network in general. To correct for this, only cases where the power law fits the distribution with a threshold no higher than 3 will be considered (Figure 11.4).

Figure 11.3. Not every power law model is actually a good fit



Top left: the power law is a good fit with the data observed. Top right: Another model provides a better fitting alternative, in this case exponential. Bottom left: Poor fit for any of the four considered models. Bottom right: the power law is a good fit, but only above a high threshold.

Figure 11.4. Concentration over time: Power law exponents by day, among supporters



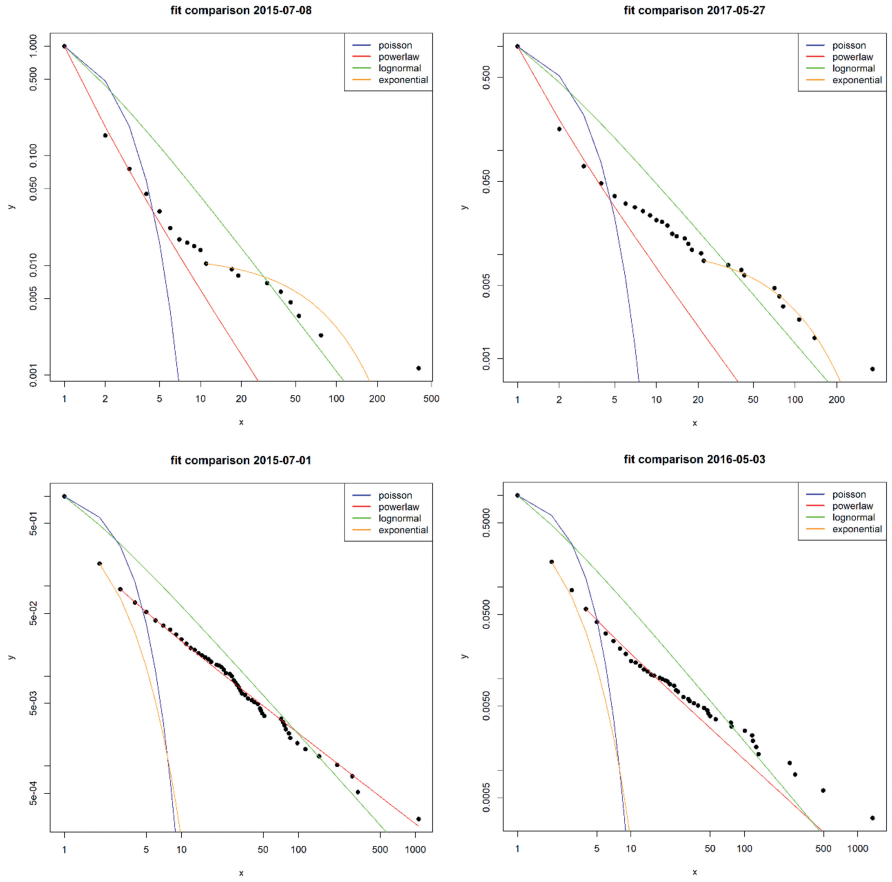
Note: Point size indicates the intensity of activity as the absolute cumulative indegree[211].

This graph is based on a calculation of the power law exponent for daily slices of the network. It reports the power law exponent among supporters, accounting for sample size, maximum lower bounds, and *p* values for goodness-of-fit. In this case, the range falls between 1.81 and 2.96. Outliers satisfy the Kolmogorov-Smirnov *p* value test; see, for instance, Figure 11.5.

The two dates with the highest exponent follow a surprisingly similar distribution, with minimal *p* values: July 8, 2015 has a *p* value of 0.33 and xmin of 1 and May 27, 2017 a *p* value of 0.15. In both cases, the power law model is a good fit with the lower part of the network, while the exponential model fits better with the core of the most active activists. In the cases with the lowest alphas, the fit of the power law model also is much better. Modeling the distribution this way allows for interpretation of how evenly attention is distributed among tweeters. The higher exponent on certain days, such as July 8, 2015 and May 27, 2017, suggests that a few individuals received relatively more attention than on other days. What stands out from these refinements is the role of leaders who receive medium amounts of attention from other Twitter users. Generally, the positions at either end of the distribution are as expected: most people receive little attention, and one or a few people receive disproportionate amounts of attention. Yet, closer inspection suggests that the movement is "more leaderful" on days when prominence on Twitter is shared on the midfield, among those receiving roughly 10 to 100 mentions.

211 Note that the graph is based on a calculation of the power law exponent for daily slices of the network. It takes into account minimum sample size, estimation of lower power law thresholds, and Kolmogorov-Smirnov p values for goodness-of-fit. Each point represents a cumulative degree distribution on a logarithmic scale: the distance between exponents of 1 and 2 is a factor of 10.

Figure 11.5. The power law model typically fits only parts of the network



Note: Model comparison for the two highest alphas: July 8, 2015 (top left), and May 27, 2017 (top right), and comparisons for the two lowest alphas: July 1, 2015 (bottom left) and May, 2016 (bottom right).

Conclusions

The adoption of computational tools allows for unprecedented ways to examine empirically the development of social movements. But these tools introduce new black boxes on an epistemological level. Reducing this dissonance between what we study (the role of social media in the development of social movements) and how we study it (digital data and computational methods) is challenging. Many seemingly inconspicuous assumptions and programming decisions potentially

have significant implications for how we interpret empirical findings. The call to the fit_power_law function is a single line of code. It invokes a package based on more sophisticated lines of code and math. By contrast, the scripts I wrote for performing the tasks described above totaled 1042 lines of code. Although this surely could be done more efficiently, the point is that each of those lines of code comprises multiple research decisions that could easily remain invisible, while having large implications for the findings produced.

These implications can, to some degree, be generalized to other studies of social movements. Widespread in network analysis, modeling power law distributions has also been adopted for studying the development of social movements (Borge-Holthoefer et al., 2011; Shirky, 2003). However, the procedure detailed above has important limitations. First, the measure ignores substantial parts of the network, by relying on lower thresholds and upper cutoffs. The estimated threshold excludes a substantial number of network nodes from the model, typically any node with a degree lower than six. In most empirical networks, this long tail represents the majority of nodes (Clauset et al., 2009). Moreover, a cutoff of the most highly connected nodes is common, for which an exponential model may be a better fit. It may be exactly those least and most connected network actors who fulfill crucial roles in the development of social movements (Barberá et al., 2015; Bastos, Piccardi, Levy, Mcroberts, and Lubell, 2018; Diani, 2003; González-Bailón et al., 2013). Practically, this means that we need to carefully qualify our claims about power law distributions in social movement networks, and critically evaluate any lower and upper bounds. In many cases, it may be more empirically accurate to impose low thresholds rather than to model a good power law fit for a fraction of the network.

Secondly, this implies that a good fit for a power law model is not necessarily evidence of a generative process of social movement development. While a power law distribution is typically associated with preferential attachment, different parts of the network may be subject to different generative mechanisms. For instance, low-degree nodes may approximate randomness, while the middle cadre obeys power law distributions, and the most highly connected nodes are subject to exponential growth. This suggests that future studies would benefit from differentiation and further qualification of generative mechanisms in social movements, necessitating qualitative inquiry.

Thirdly, the measure by itself does not provide absolute comparability. The many implicit technicalities of the procedure make outcomes susceptible to

individual technical skills, which are not readily evaluated in double-blind peer review processes, so that a high power law exponent cannot by itself be compared adequately across studies. Social movement scholars may find it more efficacious to focus on dynamic processes, and accordingly, on changes in the power law measurement over time. While this approach is still susceptible to biases, these would lessen in longitudinal studies and triangulation.

Although I have focused in this chapter specifically on the calculation of power laws, thus bypassing other approaches and tools used in my study, I believe that my argument has broader ramifications. Computational methods are widely employed, notably topic modeling and other natural-language-processing algorithms, image-processing algorithms, and structural graph algorithms, such as community detection and exponential random graph models. Each introduces unique assumptions and pitfalls that require thorough examination. For any application, it is exceptionally difficult to unpack the intricacies of the tools we use, even when we work in multi-disciplinary teams where statistical expertise is combined with social theory.[212]

More broadly, this chapter illustrates the benefits and limitations of a mixed-method approach to the analysis of digitally networked social movements. Initially, the power law measure is deceptively straightforward: it indicates how prominence is distributed among people using a particular hashtag. This is interesting, because the supposedly open nature of social media would suggest that, in principle, everyone could receive the same amount of attention. Additionally, one might expect that users tend to pay attention to people whom they know well. Because different users know different people well, a likely outcome would be that overall attention is evenly distributed. However, it turns out that, overall, the distribution of attention is overwhelmingly unequal. Online prominence is not only non-random; *everyone's* attention is primarily directed to a select few people. These findings led to an iterative refinement of both measure and interpretation, a result for which qualitative inquiry proved indispensable. Contextualization made obvious that many of the network nodes, even some very central ones, were opponents which needed to be filtered out, given the objective of examining the movements' leadership. It further resulted in a focus on changes over time, allowing us to better understand the development of social movements. For instance, during certain periods, attention was more concentrated among a few leaders than during other

212 Highly unlikely, even: the Riemann hypothesis remains unsolved at the time of writing (http://www.claymath.org/millennium-problems/riemann-hypothesis, retrieved July 30, 2019).

periods, depending upon the topics under discussion. Based on qualitative inquiry we knew that specific leaders were consistently prominent. Thus, consolidation became a key aspect of what we sought to explain, and helped us to realize that the power law measure captured only concentration, not consolidation.

Finally, this chapter suggests practical lessons for the development of digitally networked movements. The scale-free properties of a network of activists has implications for the way information diffuses among them. Hubs, the most prominent activists in the network, fulfill an important role in spreading information among otherwise unconnected groups of activists. When information goes viral in digitally networked grassroots, these hubs play a key role by "building followers like everyone else." This makes networks vulnerable: without these specific hubs, information is less likely to be shared widely, with less diverse information being accessible throughout the network.