The Complexity of Urban Information Gathering

by

Abdulfatai Popoola

A Thesis Presented to the

Masdar Institute of Science and Technology

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in

Computing and Information Science

Abstract

Information gathering efforts are routinely organised to get information about fugitives, help people during crises or during search and rescue missions. The success of any information gathering effort is largely dependent on the environment it is carried out in: for example, search efforts in hostile environments are unlikely to succeed compared to those in friendly areas. Also, well-planned and organized cities might be easier to search compared to ill-planned or disorganized cities.

This thesis introduces a new way of measuring the difficulty of information gathering: the eigenvector centrality distribution of the dual graph of a city's road network; this measures the probability that a random walker stumbles on the information desired. Empirical analysis of these distributions provided us with a basis for comparision and identifying trends.

Results show that it is significantly easier to search and retrieve information in North American cities compared to European and Asian cities. Also the younger a city is, the less difficult it is to search. Finally, we show that city structural forms (planned, unplanned and partly planned) have no effect on the difficulty of finding information.

These results can be applied in urban planning and development, disaster response and diffusion modelling. Moreover, since potential hotspots can be easily identified, our model can be used in monitoring crime and search and

rescue missions.

This research was supported by the Government of Abu Dhabi to help fulfill the vision of the late President Sheikh Zayed Bin Sultan Al Nayhan for sustainable development and empowerment of the UAE and humankind.

Acknowledgments

I hereby acknowledge the contributions of my supervisor, Dr. Iyad Rahwan; whose tutelage, help and support encouraged me to grow and develop myself.

I am also grateful to Dr. Jacob Crandall and Dr. Zeyar Aung for being on my Research Supervisory Council.

I am indebted to Dr. Alex Rutherford, Dr. Nicolas Stevanofitch for their help and suggestions while I was working on the thesis.

I also acknowledge the great impact the Computing and Information Science faculty at MASDAR has had on me, the SCAIlab members and my colleagues too. It has been a great experience so far.

My deepest gratitude goes out to my family for their understanding and encouragement while I was away; it was difficult to be so far away however they kept me motivated and supported me.

Abdulfatai Popoola,

Masdar City, April 16, 2013.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 Introduction

Knowing the complexity of a city with respect to the difficulty of finding information or ease of navigation will come handy in a lot of scenarios: for people working in unfamiliar terrain, during relief planning sessions and for security purposes. Even though the topology of the city might have been altered by its development, crises or growth, the kernel of the city's structure will typically remain unscathed.

A lot of researchers have studied the complexity of cities and currently there are platforms that solve some of the problems of large scale information gathering, filtering and verification. However, to the best of our knowledge, there is no platform that solves or provides a model for the problem of targeted information gathering.

This thesis fills this void: it provides a model for the problem of targeted

information gathering by trying to measure the difficulty of finding a target at random in cities. This is done by measuring the likelihood of a random walker in these cities finding the desired object.

We examined the complexity of various cities around the world. Using dual graph representations of city road networks, we calculate the various metrics of these graphs and then analyse these graphs to determine if we can find intrinsic qualities that explain why some cities are more complex or difficult to search than others.

Two recent targeted information gathering competitions provided the motivation for this research: the tag challenge [46] and the red balloon challenge [40]. We seek to find out if there is some city characteristic that explains why the culprits were found in some cities and not in others.

Our results eventually corroborate our assumptions about the data as we show that these cities are quite difficult to search. We also show that the age of a city and its location might determine how easy it is to search: North American cities are typically easier to search than European and Asian cities while younger cities tend to be easy to search too.

1.2 Thesis Structure

This thesis is structured as follows: chapter one provides an introduction, chapter two describes the challenges of information gathering and the motivation for our study while chapter three provides a literature review of related work and complexity measures. We present our methods in chapter four: our data, processing framework, approach and processed data characteristics. Finally, chapters five and six describe our experimental results and provide the conclusion respectively.

CHAPTER 2

Background

2.1 Introduction

Disasters, man-made and natural, cause large numbers of casualties, damage infrastructure and displace large populations. In recent times, tsunami, hurricanes, riots, protests, floods and acts of terror have all caused some degree of discomfort or disrupted normal life patterns in several communities. The measured impacts of crises are severe: in 2011, natural disasters caused more than 30000 fatalities, created over 240 million victims and led to economic damages worth 366 billion US dollars [16]. In 2010, the estimated loss due to natural disasters was about 120 billion US dollars [17].

Crises can change the topography of a region and induce huge population displacements as affected individuals seek refuge; consequently, existing maps and data are rendered obsolete. Thus, one of the most critical needs of relief agencies during and after crises is accurate information: rapid access

to reliable estimates of casualties and needs, damage to infrastructure (such as roads, power and medical services) and security evaluations of affected areas is essential. Such information is required in planning relief efforts, distributing resources (aid packages and rescue teams) and estimating resource requirements [18, 10].

However, information gathering processes in disaster management and recovery policies have three flaws: they ignore local informants in information gathering and situation evaluation, rely on inefficient and extended relays of information from a few trusted sources and focus on the use of overstretched and possibly damaged emergency facilities [18].

Information-gathering efforts by aid organizations usually employ approaches such as questionnaires, interviews and field deployments; eye-witness reports are usually deemed untrustworthy by relief organizations unless verified by known experts [56]. This disregard for unsolicited live reports coupled with the overt reliance on a few - and possibly overburdened - experts leads to situations of outdated or incomplete situational awareness.

Since most crises are emergencies which require near-immediate action, agencies are in a dilemma: retrieving information from crises' sites might cause delays which might be too costly (e.g. loss of lives, economic damage) while acting "blindly" based on unconfirmed information (which might eventually turn out to be false, selfish or even malicious) is also undesirable.

A unifying response common to all disasters is the establishment of relief campaigns and efforts. These typically attempt to alleviate the sufferings of the affected populace, limit further damage to humans, animals and environment and help them to become re-integrated into their normal lives after the crisis is spent. These campaigns usually have broadly similar informational needs and face the same challenges.

2.2 Informational Tasks During Crises

This section provides a broad overview of the informational tasks that have to be carried out during crises, their associated challenges and existing solutions.

2.2.1 Large-scale Information Gathering

The proliferation of mobile Internet-ready devices has changed the dynamics of modern-day interactions, communication and information diffusion. Social networks, which have now become an essential part of our daily lives, enhance information dynamics by connecting people. Online activity on these platforms span business meetings, re-unions, hanging out with friends, fund-raising and humanitarian efforts. Also, most users of these social networks are comfortable with sharing their views and opinions or those of close friends online.

These developments have led to the emergence and growth of communities that effectively share information and have established regulatory mechanisms (which might be explicitly specified or implicit). During important events, these communities serve as channels for the rapid propagation and diffusion of information by leveraging social ties and the small diameter of the entire network; members pool information from a variety of sources and spread new information to close acquaintances.

This rapid propagation of information makes it possible to get near real-time reports of events regardless of the location or state of the affected people (people living in remote areas and/or urban places with severely-crippled infrastructure due to natural disasters or censorship). This capability is extremely useful and can be applied to a number of scenarios (e.g. emergencies) requiring quick responses and decision making based on noisy

crowd-generated and not-yet-verified information.

Palen et al. in [37] state that information communication technology (ICT) tools can be used during crises to recruit volunteers, verify claims and broadcast information. Palen et al. [38] also studied online activity with respect to the 2007 Virginia Tech massacre; they found out that crowd-sourced information was 100% accurate and was available long before the official statement by the school.

Online communities have been used to successfully used to track and monitor disasters like wildfires [60], hurricanes [25], floods [59], earthquakes [50, 11] and epidemics [30]. Twitter activity data has also been used to monitor US political conventions [25] and emerging trends [32]. The potential of these platforms for information generation is amazing: during the 2011 Virginia Earthquake, tweets were posted at a rate of 5500 tweets per second, similarly there are 4.7 million tweets related to the Chilean earthquake [19].

"Live" reports generated by eye-witnesses and bystanders are of immense value during crises; they can be used by relief organizations who need to rapidly plan and coordinate relief efforts. They also provide a means for affected people to reach out to their loved ones as well as serve as "unverified" information sources to traditional media institutions.

Ushahidi[1] (a Swahili word for testimony) is a popular and successful disaster response platform based on crowd-generated reports. This open-source platform has been used to monitor and collect information about elections, civil unrest, disease tracking and emergency response in a lot of countries [15]. It also incorporates visualization tools which aid decision making.

During the 2010 Haitian Earthquake of 2010, volunteers worked assid-

[1] www.ushahidi.com

uously to create an accurate map of the stricken areas on openstreetmap[2] in less than 48 hours. This large-scale information gathering effort was so successful that a lot of aid and relief agencies switched to the maps created by volunteers [21].

2.2.2 Information Credibility

Credibility ratings are directly dependent on how the informed person perceives and judges the information he is receiving; although some characteristics of this information might influence the reader's opinion and judgement. For example, information from proven media sources are typically held to be credible due to the strict and rigorous editorial processes employed by such organizations; moreover, most media organizations are household names and are accepted as authentic sources by a large number of people.

Credibility Models

The Fogg's prominence-interpretation theory is a credibility model that attempts to explain how people assess the credibility of websites [13]. Prominence is determined by the visibility of online content while interpretation is related to how users evaluate and understand online content. Fogg also listed a couple of factors including motivation, skill levels, context, culture and browsing environment as influencing the user's perception of credibility.

In [48], Ratkiewicz et al. posited that the spread of rumours in social networks differs from Rapoport's viral model [47] of infectious disease propagation. According to [48], the plausibility of a rumour to a node is proportional to the number of its neighbours who believe the rumour is true. Once the number of infected nodes in a network exceeds a certain threshold,

[2] www.openstreetmap.org

the entire network is compromised and the rumour is seen as credible/true. As such, rumours might eventually turn into "accepted facts" by virtue of their propagation in networks. This model explains the notion that popular belief is usually accepted as being true.

Online Information Credibility

People are more likely to view information that is widespread or from an acquaintance as true and subsequently re-broadcast it within their own networks [26] (i.e. inform their connections, thereby setting up new information cascades). Accordingly, once a certain percentage of the connected population is compromised and tricked into believing false news; misinformation campaigns can become difficult to stop due to the diffusion of the information by unwitting participants.

Self-interested and malicious agents can thus exploit the information diffusion characteristics of social networks to successfully subvert such information gathering efforts; this can lead to misinformation campaigns, the spread of rumours or smears and compromise highly-needed aid efforts in crises scenarios. Such attacks, if properly coordinated using the appropriate techniques (targeting, viral spread, proper phrasing etc.), can successfully change the perception of the public about some event. For example; in 2009, nine fake accounts were used to initiate a successful smear campaign against one of the Massachusetts senatorial aspirants [34].

Earle [11], also proved that the same properties of Twitter that make it easy for users to spread news about currently-happening disasters also make it easy to spread false stories and rumours and misinform people. Mendoza et al. [33] in their work on the Chilean earthquake highlighted the challenges of using community-sourced information. They found out that reports were

rife with rumours and misinformation; however, false information came under more scrutiny by users than true reports.

To combat these issues, most community sites now employ corrective mechanisms such as moderation and flagging to cope with the high variance in online media accuracy. As such, the quality of content curated on these sites tends to improve over time as errors and inaccuracies are removed.

The October 2007 wildfires in Southern California provide an example of information verification; administrators of the community site rimoftheworld[3] collaborated with local authorities in disseminating and verifying information. Furthermore, they conducted physical investigations of damage caused by the fires and shared it online [55].

2.2.3 Coordination and Task Distribution

Crises cause distress in people - they can bring out altruistic behaviour in people or even worse behaviours such as looting, mob actions etc. Once the initial shock is over, affected people have to fix their disrupted life patterns or live with the effects of the crisis.

The use of social media and the Internet is not limited to information retrieval alone; it has also made it possible for volunteers who are thousands of miles away from disaster epicentres to contribute and help in relief efforts. Whereas physical recruitment is constrained by a number of limits, online participation is virtually limitless. Such volunteers assist with recruiting other volunteers, translating messages and helping with requests [21].

Studies of humans under crisis show that they are typically level-headed and work towards helping one another and recovering from the disaster. In the immediate aftermath of crises, human groups form amongst the affected

[3] www.rimoftheworld.net

populace with the aim of providing help to seriously-wounded, evacuating feeble members and helping to provide security [45]. Thus, it is possible to employ these citizens in recovery efforts or task distribution during such situations.

In [44], the authors reported the results of an online survey carried out on a population of people displaced by Hurricane Katrina. They showed that the affected people sought to establish communities online and used these communities to find emotional support by connecting to people who had been through the same ordeal. Other online support communities of the same dedicated to the distribution of aid also emerged [57].

The Sahana Foundation[4] produces software platforms for this purpose; their open-source software offers support for tracking and categorizing affected people, monitoring their needs and matching donors to requests.

2.2.4 Information Filtering and Data Extraction

With the ubiquitousness of mobile devices and Internet access, it is easy for get live reports based on the observations of people in affected areas. This information is then shared and distributed over the Internet. A shortcoming is the difficulty in making sense of such huge streams of information from multiple sources. Most reports come in a wide variety of formats (including structured and unstructured information); this makes filtering for specific information or synergizing content challenging.

Although traditional media such as newspapers and TV stations provide general awareness about events as they happen and can provide structured easy-to-use data, their reports are sometimes sensationalized, exaggerated and not detailed enough to plan recovery efforts [53]. They might also be

[4] www.sahanafoundation.org

biased towards more popular areas or items, ignore relatively undeveloped areas or miss out on certain areas in their coverage [55]. Official repositories of information which provide extensive coverage suffer from slow updates; disaster scenarios render these sources outdated making them useless to information seekers and crisis-stricken people.

The crowdsourced disaster response team that used crowdflower[5] provides an example for this category. The team was able to create a flexible and easily scalable platform which was successfully used in translating, filtering and geo-tagging messages originating from Haiti; a volunteer task force made up of people from all over the world provided the manpower needed [22].

2.2.5 Targeted Information Gathering

New behavioural patterns emerge during crises - volunteer relief efforts by compassionate people, collaborative efforts to find survivors or compile lists of affected people and locations, adoption of new technology and new communication and interaction patterns.

Thus, every new crisis brings along a new digital trace of information available in the form of text messages, emails, blog posts, tweets, online communities, social groups and other custom fora.

2.2.6 Challenges of Information Gathering

The main challenges regarding information generation during crises mostly revolve around trust, accuracy and reliability. Other minor issues include quality and objectivity of reports, misinformation by malicious agents, handling huge rates of information flow and language barriers.

[5] www.crowdflower.com

Another issue is the absence of a central repository to which user generated reports can be submitted; such a pool which should have independent verification and filtering capabilities will provide the disparate aid organizations with data. This will eliminate the duplication of efforts and aid deployments.

Table 2.1 provides a summary of the various categories, challenges and existing solutions for tasks in crisis response.

2.3 Motivation

This study traces its roots to the tag challenge [46] which provides an ideal example of a targeted information gathering scenario. The winning strategy involved a recursive reward mechanism that incentivized inhabitants of those cities to look out for the five "culprits". Despite the appeal of monetary reward, only three culprits were found by all participating teams.

Consequently, we seek to understand why the culprits were found in certain cities and not discovered in others; we posit cities share common latent characteristics that explain this phenomenon. The difficulty of finding a person in a city depends heavily on the complexity of the city's road network, this in turn determines how much information is needed to search the city.

Also, to the best of our knowledge, the targeted information collection category is the only informational task category without any readily available solution; all other categories have platforms that attempt to solve or mitigate the challenges. We assume that this is probably due to the challenge of modelling the situations involved.

Thus, by analyzing information networks of various cities including the cities of the tag challenge, we aim to model their complexities and identify

	Large scale information gathering	Information filtering and verification	Coordination and task distribution	Targeted Information Gathering
Description	Mapping hit areas. Crowd-sourcing Information. Gathering information about events. Citizen Journalism.	Extracting information from lots of noisy data. Filtering repositories based on credibility	Assigning tasks to volunteers. Monitoring logistics. Integrating feedback from field reports in planning	Specific requests such as the location of a point or information about the whereabouts of a person. General Questions about some event or place.
Challenges	Volunteers. Access to the Internet. Collection of Information in transit. Retrieving information from remote areas, conflict zones, censored regions and places with damaged infrastructure.	Availability of data, computational power and methods. Information obsolescence and redundancy. Unstructured and widely-varying information formats. Communication vis-a-vis cultural differences and language barriers.	Access to accurate information, Misinformation and sabotage	Urban complexity.
Existing Solutions	Ushahidi, Twitter, OpenStreetMap, Smartphone apps	Swiftriver, Storyful	Sahana, Crowdflower, Mechanical Turk	???

Table 2.1: Informational Tasks in Crisis Response

trends and patterns based on interactions between their components.

Our approach differs from existing work based on information networks of cities; not only do we do a large scale analysis of a lot of cities, we also factor in the effects of other variables that characterize cities such as its population, age, location. This, we believe, gives a more holistic image of the city's culture and fabric.

CHAPTER 3

Measuring Urban Information Gathering Complexity

3.1 Introduction

Cities form an integral part of human history and continuously sustain the development of new ideas and efforts. In early times, human settlements were simple and scattered; however cities have grown and become more complex.

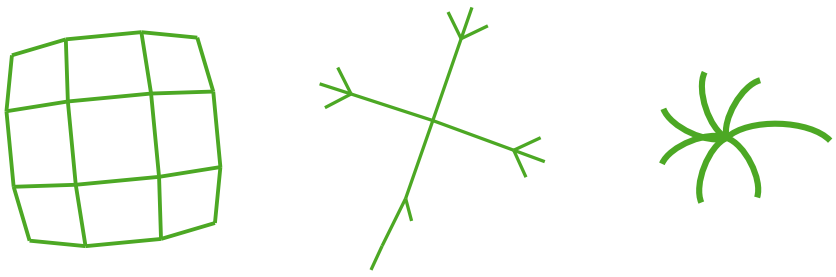
The growth of cities can be viewed as an evolutionary process: settlers cluster around some desirable location e.g. an oasis, a port, a stop on a trade route or a refuge. Over time, more settlers arrive leading to the expansion of the settlement and the creation of more buildings, roads and infrastructure. This cycle is then repeated indefinitely.

The growth and development of cities might be planned or unplanned: some cities are designed to have highly-regular patterns while others grow haphazardly. In planned cities, it is possible to describe the network pat-

tern (it could be a grid network or a series of concentric circles); on the other hand, there is usually no way of describing the intrinsic patterns of unplanned cities. Fig. 3.1 shows some layout patterns.

Cities are complex entities that encompass interactions amongst multiple interacting agents with varying needs. The factors influencing the complexity of a city include its size, population, age and shape; other factors such as the city's type, location, inhabitants' culture etc. also contribute to the overall complexity. Emerging needs such as urban area planning, social and demographic requirements, security measures, mobility patterns, accessibility measures and resource distribution dictate a need to fully understand the dynamics of cities and the factors influencing them.

Since city networks are complex networks [35], it is possible to apply graph theoretical models and principles from network science methods in analysing and solving city challenges. The use of concepts from network theory provides researchers with robust tools and a platform for evaluating cities with respect to their complexity and structure.



(a) Grid layout (b) Hub layout (c) Star/Spoke layout

Figure 3.1: Sample City layouts

3.2 Related Work

Complexity studies of cities cover a diverse range of fields including but not limited to city representation models, analysis of network properties, complexity evaluations, algorithmic proofs of representation formats and complexity evaluations. The various categories are described in the following sections.

3.2.1 City Network Analysis

These studies measure the various characteristics of city networks to detect latent relationships between nodes and to characterize cities exhibiting certain properties.

Crucitti et al. [8] did a thorough evaluation of 5 different centrality distributions for 1-square mile samples from 18 different cities, with each city represented as a primal graph. Their results showed that some of the centrality measures (e.g. closeness, betweenness and straightness) were similar for all cities however two centrality measures had varying distributions across the 18 city samples. They ultimately found out that self-organized cities conformed to a power-law model and differed from planned cities.

Strano et al. in [54] investigated the structural and network properties of 10 primal graphs of European cities: they carried out a principal component analysis of the centrality distributions of the various cities first, and then identified clusters by measuring the Gini coefficients of the centrality distributions. The cities had broadly similar structural characteristics even though they had distinctive geometrical features.

Buhl et al. carried out a topological analysis of a large number of urban settlements in [7]; their study used graph-theoretic principles to analyse and study the growth, evolution and functioning of various unplanned areas.

They discovered that urban settlement patterns were efficient and more robust to failures.

Other work in this area include the detection of similar clustering patterns in cities [4, 9] and the identification of the defining characteristics of urban city networks using a primal graph representation [20].

3.2.2 Evaluations and Comparisons

Researchers in this area typically evaluate some metrics for a variety of cities or compare network metrics based on some criteria.

In [43], Porta et al. applied a novel network assessment method to the problem of urban design. Their framework, which considered spatial distances and various centrality measures on primal graphs, was successfully used to select the best option from two design scenarios.

Building on work done on centrality distributions of cities, Scellato et al. [51] identified the skeletal street networks (named backbone) that were essential to a city. City backbones are calculated by building spanning trees based on the edge betweenness and information characteristics. They extracted the backbones for two different cities and were able to show how the backbones influenced mobility patterns, crime distribution and commercial activities.

Jiang et al. [27] carried out a comprehensive topological analysis of areas derived from 40 American cities. They found out the dual graph representations of these networks using the street-continuation principle exhibited scale-free behaviour and the small-world phenomenon with respect to street length and degree. Earlier work by [28] however showed that dual graph representations based on the named street approach exhibit the small-world behaviour and not the scale-free behaviour.

Masucci et al. [31] analysed the dual and primal graph representations of London. Using a benchmark of three artificially-generated graphs, they found out that the London graph exhibits self-organizing properties.

Porta et al. in [42] found differences between heterogeneous and homogeneous cities; they also advocated the use of the primal graph representation for the network analysis of cities.

3.2.3 City Graph Representation Formats

There are two major classification formats and each has its strengths and weaknesses.

Primal Graph Representation

The simplest and probably the most intuitive way to represent road networks is the primal representation. Intersections and road end points are represented as nodes and the roads between these points are identified as edges [3, 42]. Distances between physical locations are the costs of traversing any edge in these networks. The main strengths of this approach are its wide adoption, ease of use and simplicity. However, networks represented in this format might not exhibit certain behaviours such as scale-free properties.

The primal graph can be defined as $G = (V, E)$ where $\forall i, j \in E$; $e_i e_j \in E$ if there is a road between the endpoints i and j; otherwise $i \in V$ where i is an endpoint or intersection.

Dual Graph Representation

The dual representation is based on the space syntax methodology [23] which focuses on the accessibility of a particular space from other spaces in the same network. The accessibility of a space from another space is defined

as a journey that has the fewest changes in direction or requests to seek information [24]. Space syntax has emerged as a great way to use graph-theoretic measures to quantify the complexity of cities.

This representation format is unorthodox and not as intuitive as the primal representation: roads are represented as nodes and their intersections are the edges in the network [41]; it has also been called the information city network [49].

Thus, given a road network, we define the dual graph as $G = (V, E)$ where $\forall i, j \in E$; $e_i e_j \in E$ if roads e_i and e_j intersect; otherwise $e_i \in V$ and $e_j \in V$.

A major challenge associated with using the dual graph representation is the issue of preserving street identity over long distances. The use of street names as a criteria for merging roads was used by [28] in identifying similar streets. Another method to preserve street identity is the Intersection Continuity Negotiation (ICN) [41] which collapses streets based on the angle between their edges.

Errors can be introduced while using these methods: the named-streets approach might merge unrelated streets that have the same name while the ICN approach might merge streets that are totally unrelated.

Figs. 3.2 and 3.3 show the actual and dual graph representations for a section of Abu Dhabi.

Criticism of Representation Formats

Criticism of the dual graph approach include the bias inherent in the graph creation process, the absence of a limit on the number of edges a node might have (this happens because of the subjectivity in the graph construction phase and as such some roads might have extremely high number of

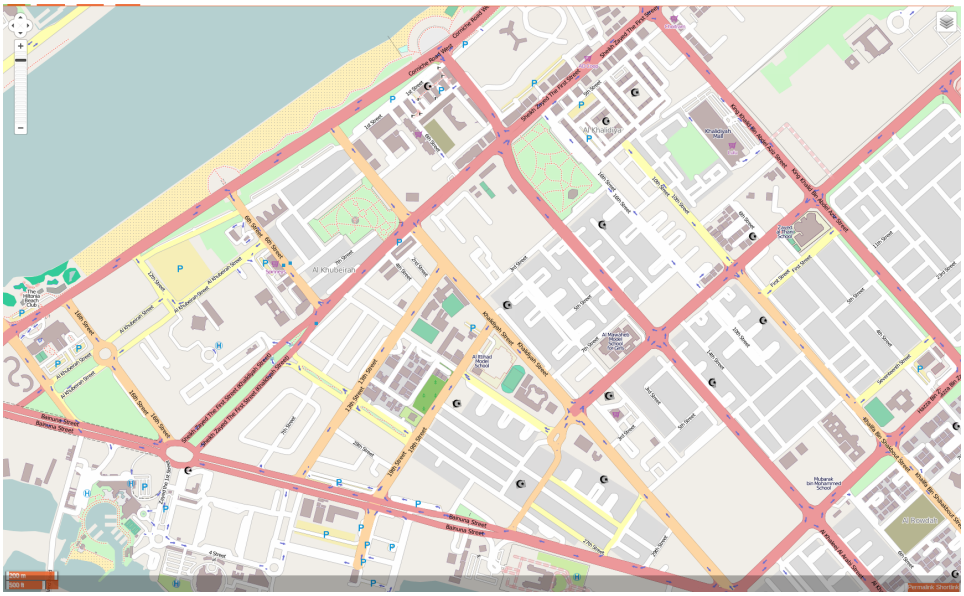


Figure 3.2: Map of a section of Abu Dhabi

edges) and the absence of distance information in such representations. However, this representation format brings out latent properties of such networks (such as scale-free behaviour) due to the possibility of having a node with a high number of edges since edge numbers are not limited, which might not be observed in primal graph representations.

3.2.4 Miscellaneous

Another interesting piece of work was done by Turner et al. [58] who defined an algorithmic framework for defining axial maps of axial spaces as defined by [23].

Omer in [36] departed from the orthodox use of centrality measures as a yardstick for graph complexity analysis and instead innovated a multi-perspective approach that combined graph theory and Q-analysis. This enabled them to factor in multi-dimensional chains of connectivity and led to the discovery of patterns and relationships between city structural qualities

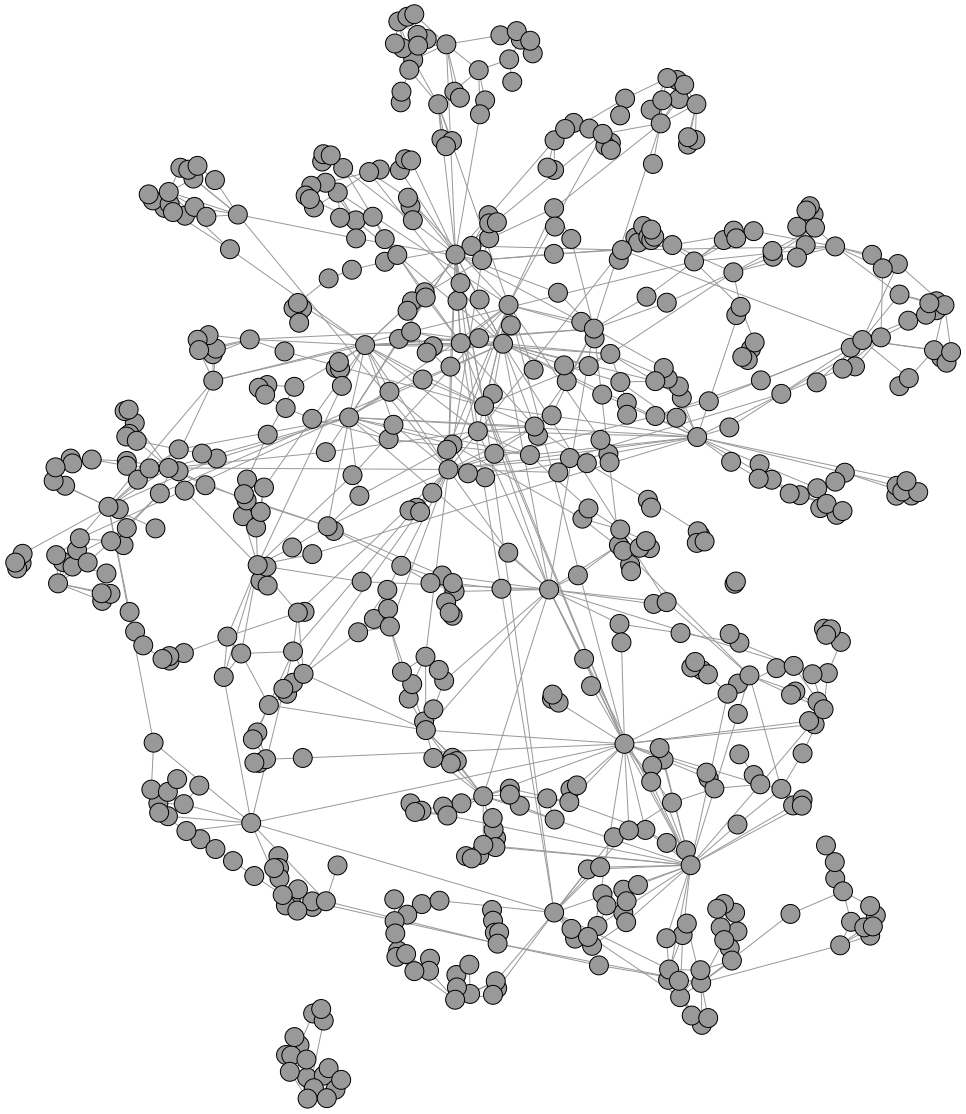


Figure 3.3: Dual Graph Representation of the same section of Abu Dhabi

and global image.

3.3 Complexity Analysis

One of the most challenging tasks in complexity analysis is determining if a particular system is complex and if it is, measuring its complexity. To compare systems, there is a need to have a yardstick. Existing yardsticks

include biased and subjective human ratings based on heuristics, intrinsic system parameters, properties or characteristics and information content.

3.3.1 Measuring Information

The amount of information that is known about a system can be used as a measure of its complexity. There are two major schools of thought regarding information theory: the first group views information availability as a measure of complexity while the second group views information as a relative measure derived by comparing the information values of several complex systems.

The first group follows the approach of Claude E. Shannon who defined entropy as a measure of information and complexity in his seminal paper [52]. The other school of thought which derives from Solomon Kullback, however, uses relative measures of information.

The major difference between both measures is the fact that the Shannon measure quantifies uncertainty while the Kullback measure is a measure of information gain [12]. However since the Kullback measure requires a prior, the Shannon measure has the added advantage that it is not affected by event ordering.

3.3.2 Entropy

Entropy, an abstract quantification of information, is a fundamental measure in information theory; the entropy of a distribution is a measure of the expected value of the information that can be derived from the distribution and this can be used in estimating the complexity of systems. Entropy values depend on the size of the system and the event probability distribution; as such, systems with a large number of possible events tend to have higher

entropy values.

Entropy, Information and Search Complexity

Assuming we want to know how much information we can gain from the probability distribution of a set of possible events. The occurrence of less-likely events (events with an exceedingly small chance of occurrence) reveals a lot of information; conversely, when likely events (events with high probabilities of occurring) happen, the information gained is small. More lucidly; if an event occurs with probability 1, then the information gained by knowing that the event occurred is 0.

Thus, the information gained by the occurrence of an event is inversely proportional to the probability of that event occurring, highly-likely events reveal lower information and low-probability events give high information. We can go ahead to state the characteristics of Information I with in terms of an event e which has a probability p of occurrence.

1. The Information associated with any event is never negative i.e.

$$I(p) \geq 0. \tag{3.1}$$

2. If an event is certain to occur i.e. the probability of its occurrence is 1, then the information gained when that event happens is zero.

$$I(1) = 0. \tag{3.2}$$

3. If independent events occur, the information gained from these events occurring is the sum of the individual information values for each event.

$$I\left(\bigcap_{i=1}^{n} P_i\right) = I\left(p_1\right) + I\left(p_2\right) + ... + I\left(p_n\right) \tag{3.3}$$

Where

$$\sum_{i=1}^{n} P_i = 1 \tag{3.4}$$

4. Information is inversely proportional to an event's probability of occurring.

$$I\left(p\right) \propto \frac{1}{p} \tag{3.5}$$

Based on the criteria listed above, the information gained when two independent events happen is inversely proportional to the product of their probabilities.

$$I\left(p_1 \cap p_2\right) \propto \frac{1}{p_1 p_2} \tag{3.6}$$

However, this value must be equal to the sum of their individual information values:

$$I\left(p_1 \cap p_2\right) = I\left(p_1\right) + I\left(p_2\right) \tag{3.7}$$

The only mathematical function that satisfies all these criteria is the logarithmic function. Thus, we can calculate the information gained when an event e with a probability p of occurrence happens as:

$$\begin{aligned} I\left(p\right) &= \log(1/p) \\ &= -\log(p) \end{aligned} \tag{3.8}$$

The entropy of a system is the information associated with the distribution of its values, this is the expected value of information for each event in

that distribution. Entropy, H, can thus be mathematically specified as:

$$\begin{aligned} H(X) &= E[I(X)] \\ &= \sum_{i=1}^{n} P(x_i) I(x_i) \\ &= -\sum_{i=1}^{n} P(x_i) \log P(x_i) \end{aligned} \tag{3.9}$$

3.4 Graph Centrality Measures

Graph properties that are typically used in graph analysis include the average node degree, the diameter of the network, average path distance, shortest path length, clustering and the various centrality measures [41].

Centrality measures are used to estimate the importance of nodes or the influence nodes exert on one another; the concept first appeared as a measure of relative importance in Bavelas seminal paper [5]. However, over the years, refinements have been made and this concept has been extended to other fields. Centrality measures now take into consideration the whole of range of values in the community and not just individual values [43].

3.4.1 Degree Centrality

This measures how well connected a node is in the network and is the number of the other nodes that the node is connected to. High-degree nodes in a city road network representation are typically arteries or very long roads with a lot of intersections.

The degree centrality of a node in a graph is its degree while the degree centralization of a graph is a measure in the variation of centrality values for each node. The degree centralization of a graph **G** containing **N** nodes with respect to its constituent node degrees can be calculated using the general centralization formula defined by Freeman in [14] as:

$$C_D(G) = \frac{\sum_{i=1}^{n}[C_D(n_{max}) - C_D(n_i)]}{(N-1)(N-2)} \tag{3.10}$$

where $C_D(n_{max})$ is the highest node centrality value for the Graph **G**.

3.4.2 Closeness centrality

The closeness centrality of a node is an estimate of how close the node is to other nodes along the shortest paths possible in the network.

For a graph containing **N** nodes, the closeness centrality of a node n is given as: [61]

$$C_C(n) = \frac{N-1}{\sum_{m \in G; m \neq n} d_{mn}} \tag{3.11}$$

where d_{mn} is the shortest path between nodes m and n.

3.4.3 Betweenness Centrality

The betweenness centrality of a node is the ratio of all shortest paths that pass through that node to the total number of possible shortest paths in the network. It measures the relative importance of any node in the network with regards to information flow. Nodes with high betweenness are essential nodes in any network as they link up disparate segments of the network.

The betweenness centrality of node n in a graph G is given as: [61]

$$C_B(n) = \frac{\sum_{j,k \in G; n \neq j \neq k} |d_{jk}(n)| / |d_{jk}|}{(N-1)(N-2)} \tag{3.12}$$

where

$|d_{jk}|$ is the number of shortest paths between nodes j and k

$|d_{jk}(n)|$ is the number of shortest paths between nodes j and k that

contains node n.

3.4.4 Eigenvector Centrality

This is a measure of the influence a particular node has in the network or the importance associated with it. It is based on the premise that being in a cluster of highly-important nodes or being close to such clusters will lead to higher influence values and vice versa.

This is calculated as: [6]

$$C_E(n) = \frac{\sum_{m \in G} M_{mn} e_m}{\lambda} \tag{3.13}$$

where $\mathbf{M}$ is the adjacency matrix representation of the graph G.

This can also be written in matrix notation as:

$$n\lambda = Mn \tag{3.14}$$

where n is the eigenvector of $\mathbf{M}$ and its associated eigenvalue is λ.

CHAPTER 4

Methods

4.1 Introduction

This chapter describes the experimental dataset, processing framework, theoretical approach and characteristics of the processed data.

4.2 Experimental Data

4.2.1 City Road Network Dataset

The city road network data used in this study were gotten from openstreetmap[1] XML exports. Openstreetmap is an online crowdsourced platform that allows volunteers to edit and update a map of the world with information such as roads, streets and places.

The XML exports describe geographical areas using three fields: Nodes,

[1] http://openstreetmap.org

Ways and Relations. Nodes are points on the map and contain geographical information such as latitude, longitude and tags for extraneous data; Ways represent roads or buildings; a way is a collection of all nodes that make up a road or fall within the boundaries of a building, ways also have a tag field for extra information like street names or type; lastly, Relations, which are a collection of nodes and ways, are used to describe relationships between entities.

The openstreetmap dataset suffers from the inclusion of poorly-labelled streets in export files, inconsistent street naming conventions and extensive segmentation of streets.

4.2.2 Data Processing

The creation of Way and Node object representations enabled the parsing of the XML dumps into memory. However, this simple representation was noisy and contained unneeded information; thus it had to be converted into a useful and accurate graph model.

The first task was to combine the various parts of segmented roads using the named-street approach [28]. The major shortcoming of the named-street approach is that it can merge streets which share the same name but are in different places. To eliminate the potential mismatch of unrelated streets, our method only merges intersecting streets which have the same name - intersecting roads are defined as roads which have at least a geographical coordinate in common - this approach leads to accurate merges once both conditions were met.

The XML exports for cities usually included locations outside the cities. We retrieved the geographical coordinates for each city by making requests to the Google Maps API and used these coordinates to build bounding

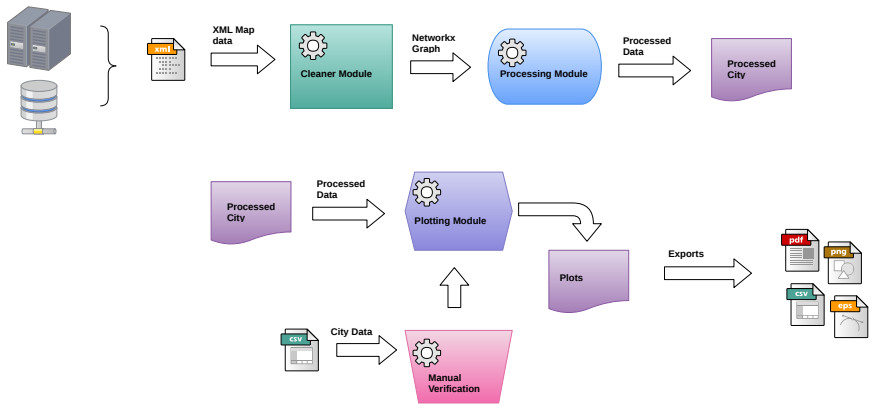


Figure 4.1: Architecture of City Processing Framework

boxes. The resulting city XML data (after the merge of segmented roads) was further refined by filtering out roads and locations that were outside the bounding boxes.

A last round of filtering was carried out to remove unwanted types (e.g. buildings) before undirected dual graph [41, 49] networks were generated. These dual graphs were built by converting the collapsed roads to nodes, finding intersections based on shared geo-spatial points between roads and then creating edges based on the intersections.

The resulting graphs were filtered to remove isolated nodes, analysed for scale-free characteristics and then passed on to the analysis framework.

4.3 Processing Framework

This section describes the architecture of the custom software framework built for the experimental analysis. The framework's modular design and data exchange format makes it easy to extend. Fig. 4.1 shows a high level overview of the platform.

4.3.1 Modules

Cleaner Module

The cleaner module consists of tools and utility functions to process the XML representation of each city. Each XML graph of the city is parsed into a city object in Python.

This module first collapses all intersecting roads in the XML file that have the same name into a single road. The iterative merge process is repeated until there are no more segmented roads in the XML dump; this occurs when the number of streets in the dump converges and does not change on successive collapses.

The next process is the filtering stage where all the outlying nodes which fall outside the boundaries of the cities are removed; all extraneous information is also removed.

The results of the merge and filter operations are used to create a dual graph representation of the city network; this is done by creating nodes for each road and creating edges when two roads intersect. The resulting graph is output from this module.

Core Module

Graphs gotten from the cleaner module are passed to the core module and it removes isolated nodes i.e. nodes with a degree of zero. The following metrics are then calculated for the resulting graph: the number of edges, number of nodes, graph clustering, total degree, average degree, size of the giant component, Freeman centralization and the number of connected components.

The graph's eigenvector centrality distribution is calculated using the

networkx[2] function. The function employs a power iteration model with an iteration threshold of 1000, cities that do not converge are ignored.

The entropy, Gini coefficient, coefficient of variation of the eigenvector distribution are then calculated and stored. To remove bias and generate a uniform basis for evaluating the various cities, a completely connected graph was used to normalized the city metrics. In the completely connected graph, all nodes have the same eigenvector centralities.

Plotter Module

The plotter module contains functions for fitting power law models to city degree distributions as well as for generating scatter plots, logarithmic/semi-logarithmic scaled plots and statistical plots. It is also flexible enough to create clusters based on user-supplied information.

Framework Module

This automates the entire process by turning it into a batch processing operation. Users specify the directory containing the OSM files, configuration options (e.g. what plots are to be generated) and graphing information. The framework reads these values and generates outputs when the process is completed.

Utilities Module

This contains a variety of helper functions that help with the following tasks: CSV conversion, OSM file handling, Geographical distance calculation, City object serialization and deserialization, cluster extraction and information retrieval.

[2] http://networkx.lanl.gov

4.3.2 Processing Metrics

Initial runs of the processing framework had to be terminated after the system ran out of memory; a rewrite of the platform focused on optimizing memory and removing extraneous information. Test runs of the optimized platform used about 14GB of memory; a 50% reduction compared to earlier versions.

A full run of the platform for 148 cities graphs retrieved from the Metro Dumps[3] took about 10 hours.

4.4 Approach

We have defined a method for analysing the complexity of city road networks by modelling the roads in a city as an undirected dual graph. The roads in a city are gotten by filtering openstreetmap city XML dumps and removing all extraneous information like buildings and landmarks.

Calculating the eigenvector centrality of this graph gives a distribution of the relative importance of each node based on the influence it exerts on the network. Normalizing this distribution gives a probability distribution that still mirrors the importance of each node.

Inequalities in this distribution of eigenvector centralities can be calculated by finding the entropy and Gini coefficients of such distributions. To remove the bias caused by variance in city sizes, the entropy values are normalized against fully connected city models of the same size. The fully-connected city model is a hypothetical city in which there are intersections between every node pair; it is trivial to search this city: every location is just one hop away.

[3] http://metro.teczno.com/, accessed 25th January, 2013.

These abstract models have maximum entropy values and can be used to measure the complexity and difficulty of search in the actual city. The normalized entropy is the ratio of the entropy value for the original city to that of its hypothetical fully-connected model.

Cities with high normalized entropy values will typically be well-laid out cities; any road in such cities will be roughly as important as any other road. Conversely, cities which have low normalized entropy values will be more complex and difficult to search due to the presence of remote regions which are sparsely connected to the main network.

The same applies to the Gini coefficient; cities which have a couple of really popular roads and few unpopular ones will have higher Gini coefficients compared to cities which have near-constant street popularity distributions.

The coefficient of variation (CV), which is the ratio of standard deviation to the mean, is a dimensionless property of a distribution that is robust to changes in data size. It allows for the compare the spread of values relative to the mean in different distributions.

4.4.1 Eigenvector Centrality

The networkx eigenvector_centrality function returns a unit vector (i.e. a vector with a magnitude of one); to remove the bias caused by varying vector dimensions, each vector is normalized with respect to the magnitude of its components. The sum of the components' magnitudes in the resulting normalized unit vector is one.

Given two unit vectors spanning different dimensions:

$$\hat{\mathbf{v}}_{\mathbf{1}} = a_1\hat{\mathbf{e}}_1 + b_1\hat{\mathbf{e}}_2 + ... + m_1\hat{\mathbf{e}}_m \tag{4.1}$$

and

$$\hat{\mathbf{v}}_{\mathbf{2}} = a_2\hat{\mathbf{e}}_1 + b_2\hat{\mathbf{e}}_2 + ... + n_2\hat{\mathbf{e}}_n \tag{4.2}$$

where $n \neq m$ and $\hat{\mathbf{e}}_i$ represents a unit vector for dimension i.

These vectors, despite having equal magnitudes of one; have different numbers of components. This implies that the components of $\hat{\mathbf{v}}_{\mathbf{1}}$ will have relatively higher magnitudes compared to those of $\hat{\mathbf{v}}_{\mathbf{2}}$ if $m < n$ and vice versa. This is due to the reduction in component magnitudes due to the increased number of dimensional projections.

To account for this bias, we normalized the entropy of each city against the entropy of a fully connected vector having the same dimension. The fully connected graph is the graph in which all possible pairs of nodes are connected. The eigenvector centrality distribution for such a graph is uniform and dependent on the number of nodes in the graph: every node has a centrality value equivalent to $1/N$.

The eigenvector centrality values for a fully-connected graph containing N nodes is a vector with each component having magnitude $1/N$. This vector can be represented as

$$\mathbf{v} = a_1\hat{\mathbf{e}}_1 + a_2\hat{\mathbf{e}}_2 + \cdots + a_n\hat{\mathbf{e}}_n \tag{4.3}$$

where $a_1 = a_2 = \cdots = a_n = \frac{1}{N}$.

The magnitude of this vector, $|\mathbf{v}| = \sqrt{a_1^2 + a_2^2 + \cdots + a_N^2}$. However, since each component's magnitude is the same; $|\mathbf{v}| = a\sqrt{N}$

Thus the unit vector of this matrix will be:

$$\begin{aligned} \hat{\mathbf{v}} &= \tfrac{a}{a\sqrt{N}}\hat{\mathbf{e}}_1 + \tfrac{a}{a\sqrt{N}}\hat{\mathbf{e}}_2 + \cdots + \tfrac{a}{a\sqrt{N}}\hat{\mathbf{e}}_n \\ &= \tfrac{\hat{\mathbf{e}}_1}{\sqrt{N}} + \tfrac{\hat{\mathbf{e}}_2}{\sqrt{N}} + \cdots + \tfrac{\hat{\mathbf{e}}_N}{\sqrt{N}} \end{aligned} \tag{4.4}$$

Assuming, the sum of the components' magnitudes is S, then

$$\begin{aligned} S &= \sum_{i=1}^{N} \frac{1}{\sqrt{N}} \\ &= \frac{N}{\sqrt{N}} \\ &= \sqrt{N} \end{aligned} \tag{4.5}$$

The final step which involves normalizing this vector based on S (i.e. dividing each component's magnitude by the value $\sqrt{N}$ so that all components add up to unity) gives the vector, $\hat{\mathbf{v}} = \frac{\hat{\mathbf{e}}_1}{N} + \frac{\hat{\mathbf{e}}_2}{N} + \cdots + \frac{\hat{\mathbf{e}}_N}{N}$. The entropy of this vector $H = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$ but $P(x_i) = 1/N$ where N is the number of nodes; therefore, H is:

$$\begin{aligned} H &= -\sum_{i=1}^{n} P(x_i) \log P(x_i) \\ &= -N * 1/N * \log(1/N) \\ &= -\log \frac{1}{N} \\ &= \log N \end{aligned} \tag{4.6}$$

Thus, the entropy of a unit vector that is normalized by magnitude is equivalent to the $\log N$ where N is the number of nodes in the graph. It should be noted that the division of components to normalize values does not affect the original vectors' characteristics as the resulting vectors are scalar multiples of the original underlying unit vectors. The normalized entropy values are calculated by dividing the entropy values obtained from the original city of N nodes by $\log N$.

The Gini coefficient measures the inequality in a distribution: a Gini coefficient value of 0 implies that there is no variation in the distribution while a Gini coefficient of 1 implies the maximum possible skew in the distribution with only one value having the highest property values. The Gini coefficient of a distribution is calculated as:

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \overline{x}} \tag{4.7}$$

Using this formula, we calculate the Gini coefficients for the eigenvector distributions. Gini values were not normalized because the completely connected Graphs had Gini coefficients of 0.

The coefficient-of-variation (CV) of the distribution is also calculated, it is the ratio of the standard deviation to the mean and this is a measure of how the values in the distribution are dispersed around the mean. The CV is a "dimensionless" measure as it does not depend on the number of values in the distribution - in our case, the number of streets in the city. Thus, it can be used to compare varying distributions. The CV of a distribution is given by:

$$CV = \frac{\sigma}{\mu} \tag{4.8}$$

The degree centralizations for each of the graphs is calculated using the Freeman general formula [14]; this measure was chosen because it gives a measure of the importance of the most central node in a graph relative to other nodes; as such it can be used to infer the layout of a network. Star networks have the highest possible value of one while other networks have smaller values.

Table 4.1 provides a summary of the various measures and their significance.

4.4.2 Experiments Using Artificial Data

To test our assumption that the Gini coefficient of a graph will be low when the variation in the graph's eigenvector centrality values is low; and high otherwise, we ran our process on four artificially generated graphs.

City Network Measure	Formula	Interpretation	Significance
Entropy	$H = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$	Random distributions have high entropy values while biased distributions have low entropy values	Cities with high entropy values (due to equally important streets) will be easy to search while low-entropy cities will contain hard-to-reach nooks and crannies and should be difficult to search.
Gini coefficient	$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \lvert x_i - x_j \rvert}{2n^2 \bar{x}}$	Measures the inequality present in a distribution: high values indicate a bias while low values imply broadly equal values	Cities with high Gini coefficient values will contain remote places that are less known while cities with low values should be easy to search. The Gini coefficient, in this scenario, is the opposite of the entropy measure.
Coefficient of variation (CV)	$CV = \frac{\sigma}{\mu}$	The CV measures how the values of a distribution vary about the distributions' mean.	High CV values indicate a huge variation in road importance and popularity while low CV indicate small variation in city centrality values.
Freeman degree centralization	$C_D(G) = \frac{\sum_{i=1}^{n} [C_D(n_{max}) - C_D(n_i)]}{(N-1)(N-2)}$	The Freeman degree centralization measures how much the degree of the most connected node in a network exceeds the degree of other nodes in the network.	Can be used to measure the layout of a city network, star/wheel networks have the maximum possible of 1 while a complete graph has zero.

Table 4.1: Graph measures and their significance

We generated four different graphs: a cycle graph, a grid graph, a random graph and a Barabasi-Albert graph.

Barabasi-Albert Model The Barabasi-Albert random graph generator [2] creates random scale-free graphs using a preferential attachment model. As these graphs grow larger, some nodes end up getting a higher degree.

Cycle Graph The cycle graph [39] consists of a closed chain of nodes; it is a connected graph and every node in the graph has exactly two edges.

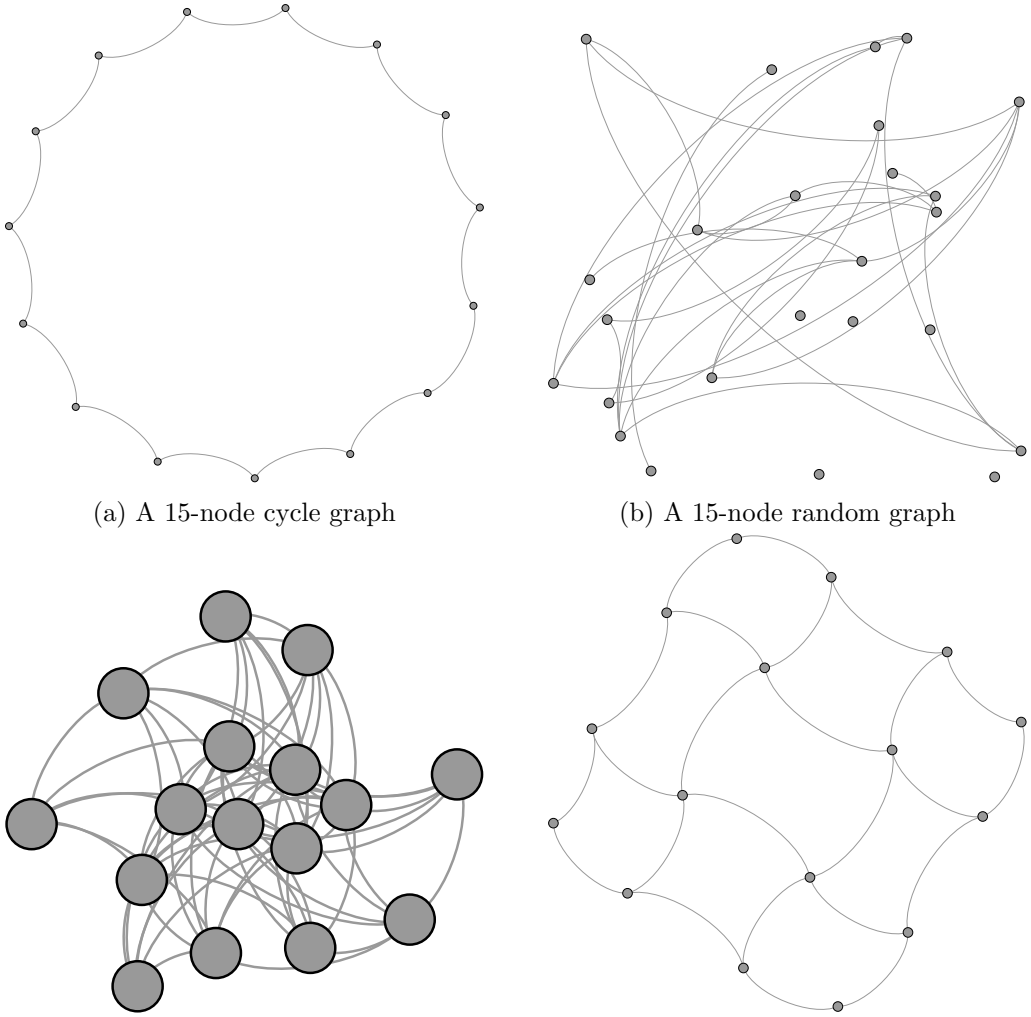
Grid Graph The Grid graph [62] can be represented as a lattice or grid and hence the name; it is obtained from the Cartesian product of two linear graphs. For our experiments, we generated only square grid graphs.

Random Graph The random graph was included as a control; for each random graph of N nodes, we chose edges in the range $[N, N^2]$. Fig. 4.2 shows visualizations of the generated graphs.

The Gini coefficient decreased in general for the graphs as the number of nodes increased. However, the Barabasi-Albert graph stood out as its Gini coefficient increased with increasing number of nodes. The Barabasi-Albert model has increasing Gini coefficient values because it is based on the principle of preferential attachment; as such some nodes end up with relatively high degrees as the graph grows larger. Fig. 4.3 shows how the Gini coefficient varies with the number of nodes.

4.5 Processed Data Characteristics

This section describes the characteristics of the processed city data.



(a) A 15-node cycle graph

(b) A 15-node random graph

(c) A 15-node Barabasi-Albert graph

(d) A 16-node grid graph

Figure 4.2: Generated Graphs Visualizations

4.5.1 Scale-Free Behaviour

The cities ($N = 148$) were analysed for scale-free behaviour, this was done by fitting power law models to the degree distribution of the roads of each city. The generated fits showed that nearly all the cities do not exhibit scale-free properties; however, this is expected: similar work by [28] showed that dual graph representation of cities using the named-street approach do not exhibit scale free characteristics.

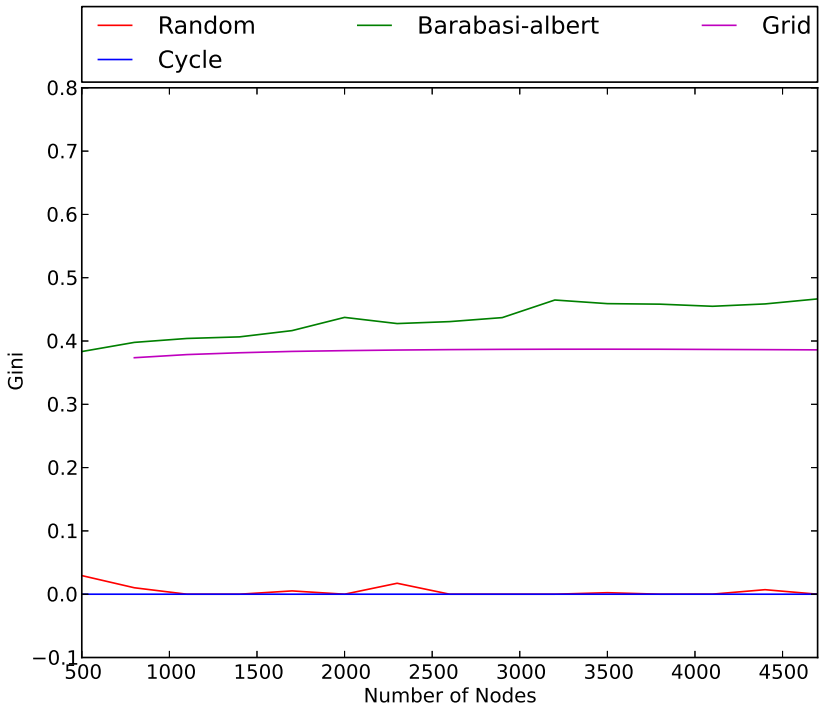


Figure 4.3: Gini vs Number of Nodes for generated graphs

4.5.2 Excluded Cities

Some cities ($N = 2$) did not converge and had no eigenvector centralities. A couple of other cities ($N = 10$) had zeros in their centrality distribution, a review of the nodes with zeros showed that they were not connected to the giant component of each graphs. All cities that fell into both categories were excluded from the experiments.

4.5.3 City Groupings

The continental and structural classes (planned, unplanned and partly planned) for 120 cities was manually retrieved from Internet sources. These information was used to segregate the cities into exclusive groups. Figs. 4.4 and 4.5 show the density plots of the cities' data based on their continental and

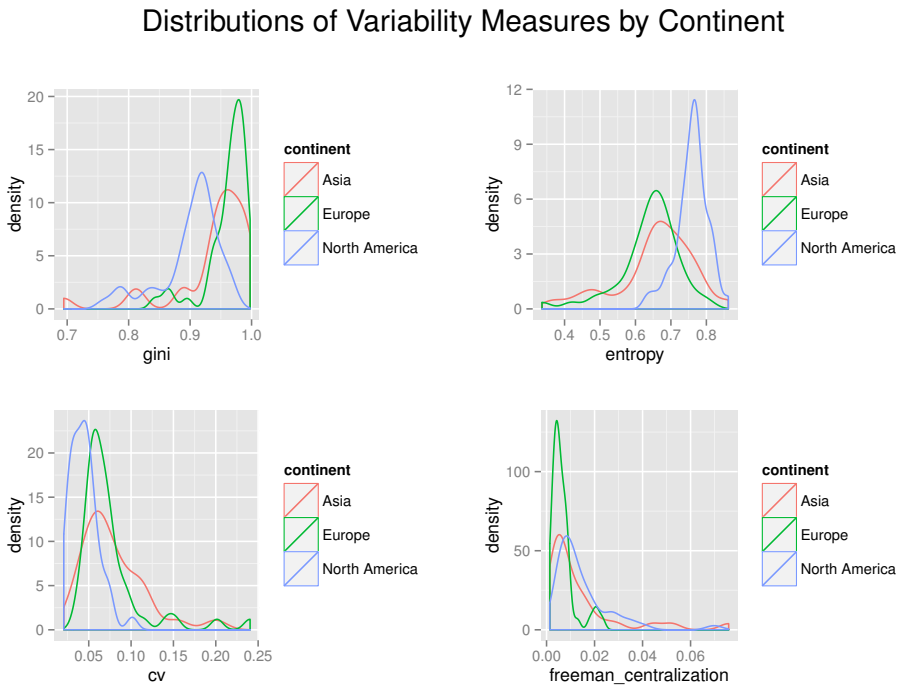


Figure 4.4: Distribution of Calculated Metrics by Continent (N=120)

structural groupings respectively.

4.5.4 Intra-Metric Correlations

Fig. 4.6 simultaneously shows the scatter plot and correlation information between all pairs of calculated metrics. The following pairs of metrics have quite strong correlations ($r(120) \geq 0.7$), (Gini coefficient, Freeman degree centralization), (Gini coefficient, entropy), (entropy, coefficient of variation) while the weakest correlation is found in the pair (coefficient of variation, Freeman degree centralization).

The high correlation between the Gini coefficient and the entropy can be explained by the fact that both are measures of inequality in distribution. For distributions with highly random data, the entropy will be high while the Gini will be low; the converse relationship also holds for the reverse case.

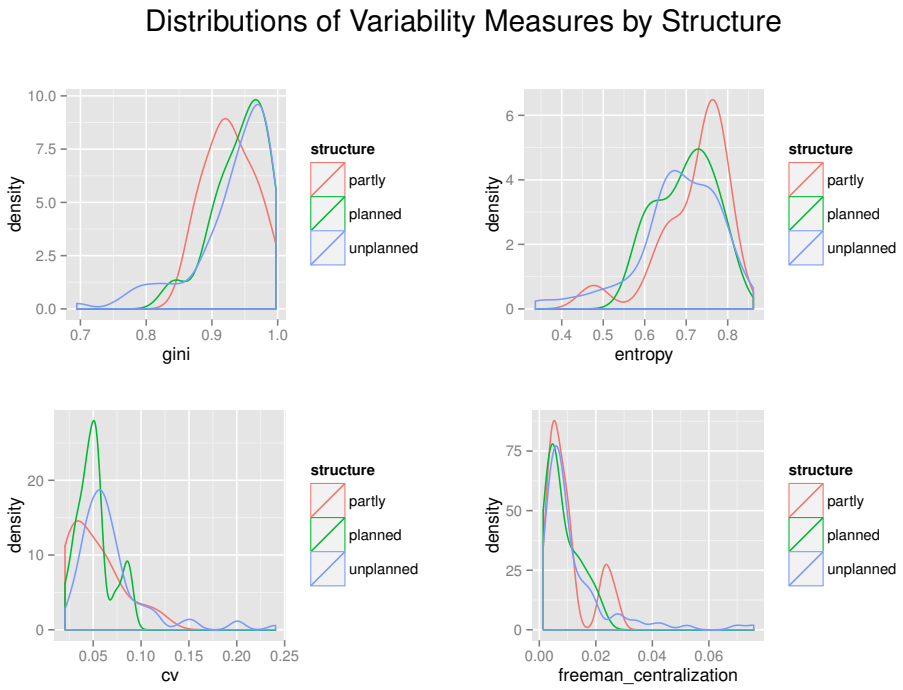


Figure 4.5: Distribution of Calculated Metrics by Structure (N=120)

Hence the reason for the strong negative correlation between both measures.

The Freeman degree centralization is a measure of how the node with the highest degree exceeds all other nodes in the network with respect to its degree. For networks having a star-like structure, this value will be very high (i.e. close to the maximum of 1). This implies that cities with high Gini coefficient values should have high Freeman degree centralization; however, the large number of roads in all the cities make these values disproportionately minute. Moreover, cities do not necessarily need to have a star-like structure to have high Gini values.

The correlation between the entropy and the coefficient of variation can also be explained thus. A high entropy means that most of the eigenvector centrality values are similar and as such the variation in values will be low. On the other hand, if the entropy is low, then some streets are more popular

and have higher centrality values relative to the others. Consequently, the distribution of eigenvector centrality values will be skewed by these larger values and this will in turn lead to higher coefficients of variation.

The low correlation observed between the Freeman degree centralization and the coefficient of variation can be explained in terms of what both metrics measure. The Freeman degree centralization is based on the degree of the most connected node in the graph while the coefficient of variation takes into consideration the dispersion of all values in the graph. As such, the coefficient of variation is much more robust to changes and this will imply that there is little or no correlation between both of them.

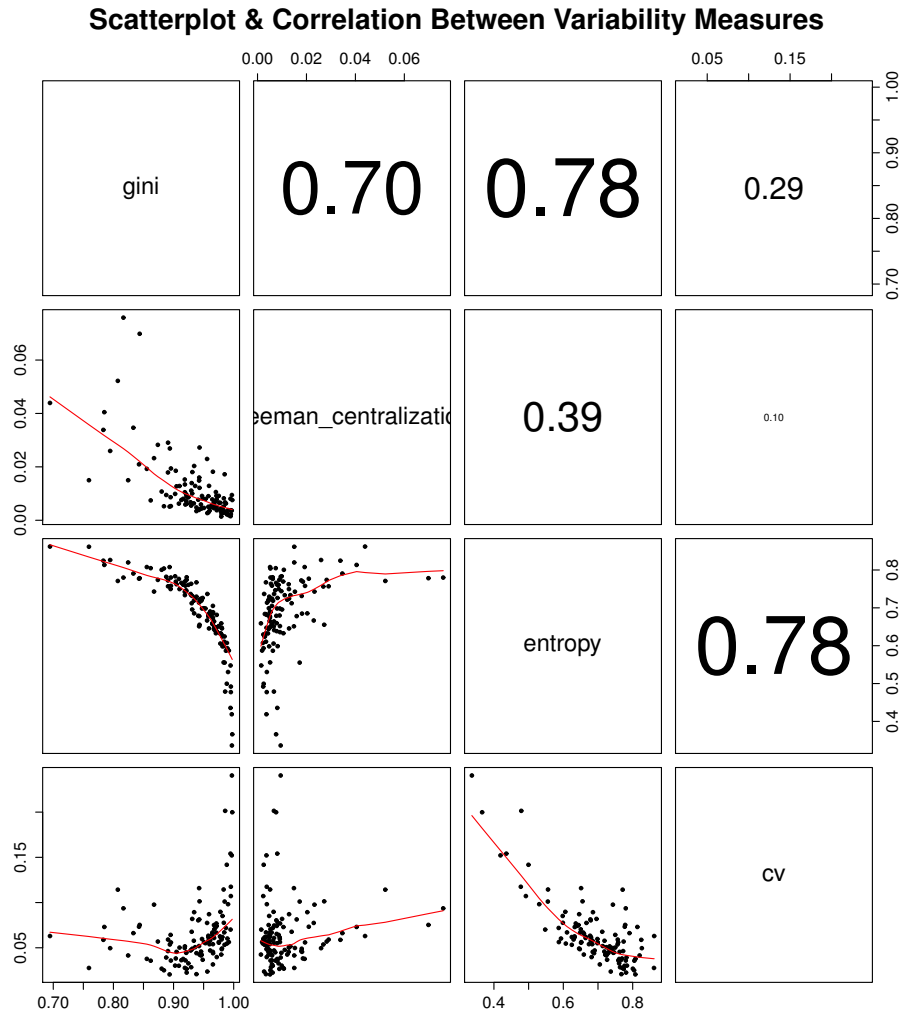


Figure 4.6: Scatter plot of network metrics showing correlation (N=120)

CHAPTER 5

Results

5.1 Experimental Results

The eigenvector centrality vector captures the likelihood of a bunch of random searchers bumping into a target located in any given street. Maximum entropy corresponds to a situation in which all search paths are equally probable, thus making it difficult for a target to hide in a street that is less frequented by people (say a small dark alley). Lower entropy, on the other hand, corresponds to a situation with far more variation in the probability of different streets being searched. It is interesting to note that both London and Stockholm, the cities in which the tag challenge winning team [46] failed to find the target, have the lowest entropy values in the histogram of entropy values 5.1.

The Gini coefficient is a measure of statistical dispersion (or inequality). It is always in the interval [0,1], making it easy to compare cities of

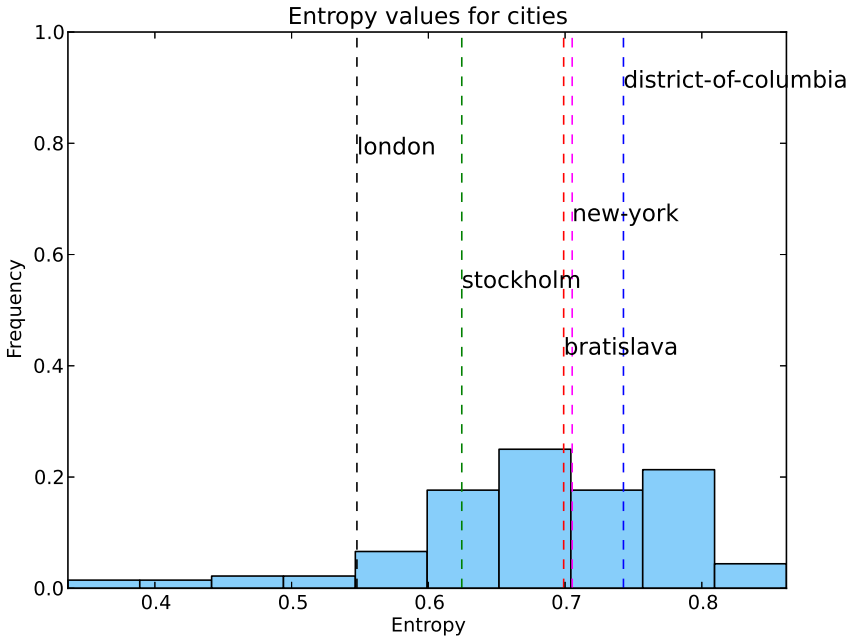


Figure 5.1: Entropy Distribution for 136 cities

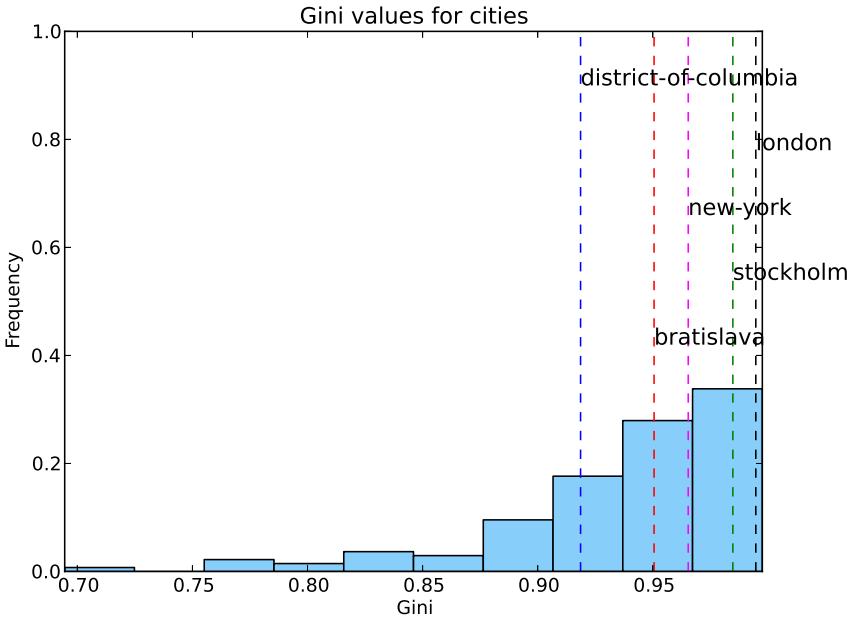


Figure 5.2: Gini Distribution for 136 cities

different sizes. In the context of urban information gathering, lower Gini coefficient corresponds to a situation with small variations in the probability of searching every street by random searchers. High Gini coefficient values correspond to a situation with high variation, in which some streets are searched particularly frequently at the expense of other streets. The observation that both London and Stockholm have higher Gini coefficients compared to the other cities of the tag challenge is also consistent with this: the Gini histogram, 5.2 shows the distribution of Gini values and the tag challenge cities.

The relatively high values of the Gini coefficient coupled with the relatively low entropy values of the eigenvector centralities of London and Stockholm provide a potential explanation for the difficulty of searching these cities. These results attribute the search difficulty to variations in centralities of different streets, making it possible to miss people unless they took highly popular routes.

Experiments on artificially generated graphs showed some relationship between the number of nodes in a graph and the entropy of its eigenvector centrality; we tried to determine if other metrics could be used to identify trends in cities. The population and area of each city were retrieved from the reported values in [1]. The founding century for each city was retrieved from data available on Wikipedia[1]. Crime data for 34 American cities were retrieved from the US census website[2].

[1] www.wikipedia.com

[2] https://www.census.gov/compendia/statab/cats/law_enforcement_courts_prisons/crimes_and_crime_rates.html

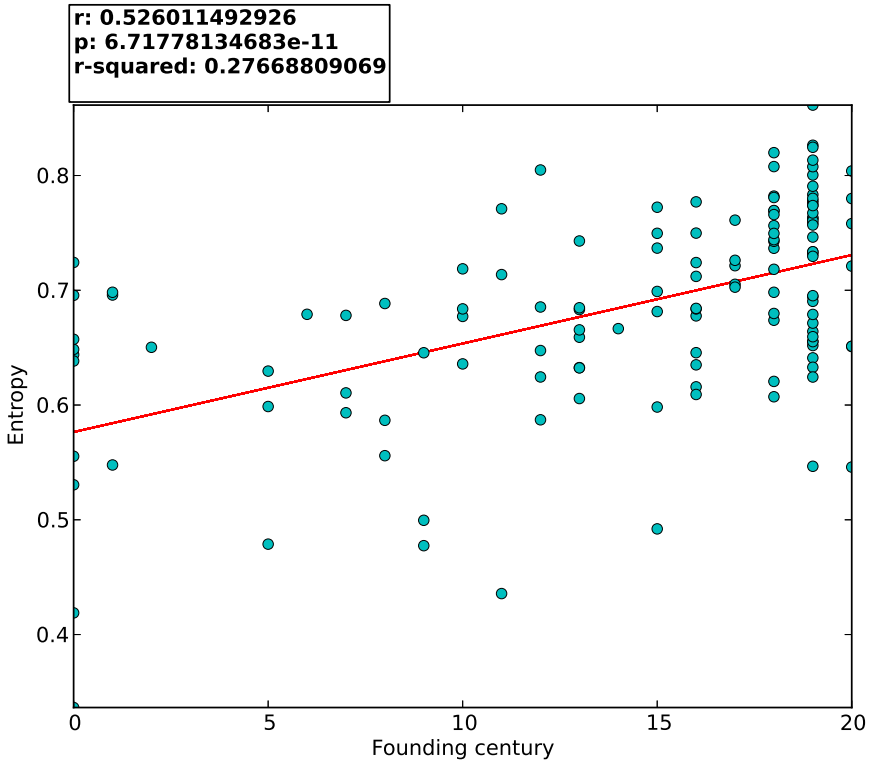


Figure 5.3: Entropy vs Founding Century for 134 cities

Entropy, Gini coefficients and Founding Century

The entropy of a city is strongly correlated to its founding century ($r(134) = 0.53, p < .05$). Similarly, the Gini coefficient of a city is negatively correlated to its founding Century, ($r(134) = -0.38, p < .05$).

The plots (Figs. 5.3 and 5.4) show that entropy generally increases as the age of the city reduces - this can be explained by the fact that newer cities are better planned than older ones. The reverse analogy also holds for the Gini coefficient as older cities are found to have high Gini values.

This explains why North American cities (which are mostly located in the USA) have higher entropies (and lower Gini coefficients) than European and Asian cities. American cities are quite younger than their counterparts in most cases and are thus mostly better planned.

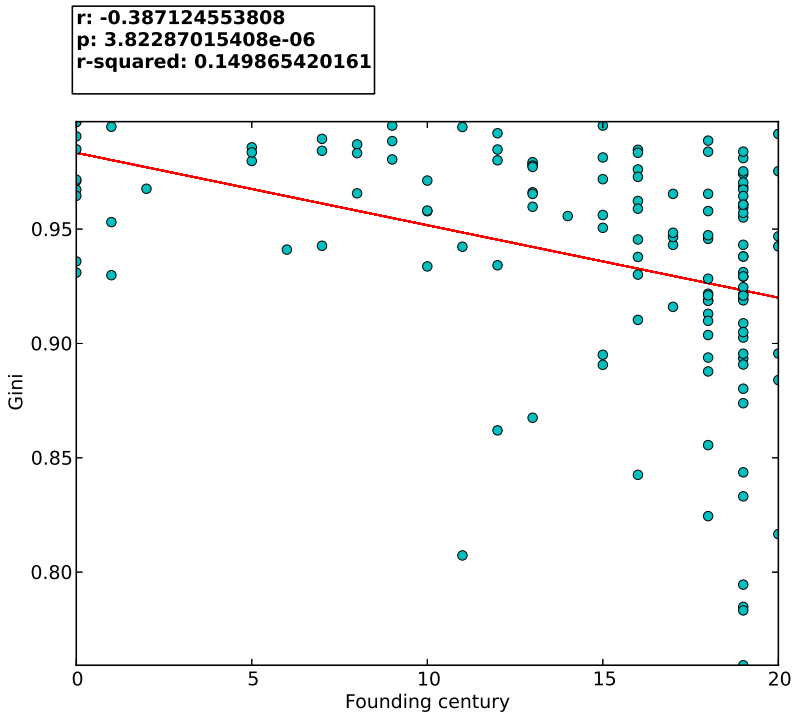


Figure 5.4: Gini coefficient vs Founding Century for 134 cities

Property Crime and Coefficient of Variation (American Cities)

Property Crime was found to be negatively correlated ($r(30) = -0.49, p < .01$) to the Coefficient of Variation, Fig. 5.5 shows the correlation.

We removed the outliers to determine how robust the measurements would be and obtained the following plot 5.6. There was still a strong (albeit smaller) negative correlation ($r(26) = -0.31, p > .05$) between the Property Crime and the Coefficient of Variation; however since the p value was greater than 0.5, the observed correlation is probably due to chance or caused by random variation in the data.

Looking at the full plots (containing the outliers), it is interesting to see that 24% of variation in the property crime rate is explained by the coefficient of variation. This behaviour can be explained thus: low coefficients of

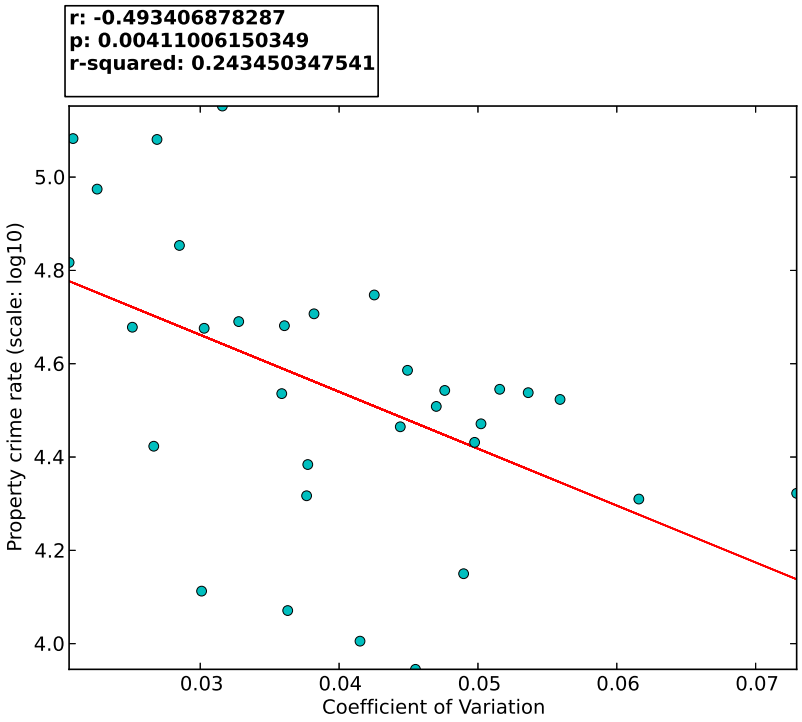


Figure 5.5: Property Crime vs Coefficient of Variation (N = 32)

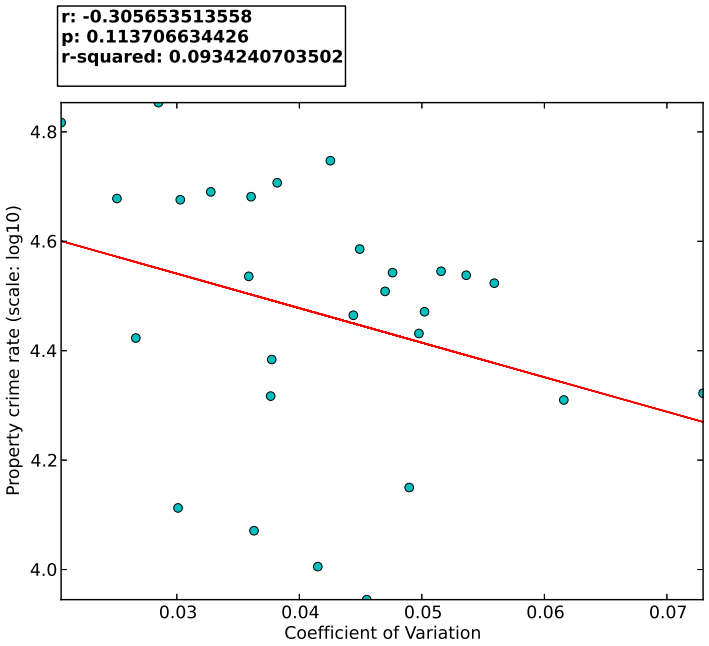


Figure 5.6: Property Crime vs Coefficient of Variation [Outliers removed] (N = 28)

variation indicate that variations in centralities are low and thus nearly all roads have similar importance while high values signify otherwise.

We speculate that the reason for high crime rates in areas with low coefficient of variation can be explained by the broken windows theory [29]. Buildings located in crime-infested areas tend to get burgled; its even easier for malicious elements to move around if all nearby roads are equally accessible. Thus, it is possible for crime to appear to 'spread' in these areas.

Other data: Population and Area

The same metrics were plotted against the populations and areas of the cities; however we could find no explicit trends in these plots.

The absence of an observed trend with respect to city areas is most probably due to the experimental approach. All experiments are based on dual graphs which ignore distances between roads; as such taking area as a predictor of the entropy (or Gini coefficient) distribution might not work. The area of a city plays a little role in determining which of its constituent roads are most important or its layout: no matter the size of a city some roads always have high thoroughfare.

The same reasoning applies to the population plots; some cities have high populations because of their economic importance or strategic locations (e.g. New York) while others have low populations. It becomes relatively difficult to find correlations between both.

5.2 Statistical Analysis

We carried out statistical tests to investigate the observed relationships between the city groupings (continental and structure).

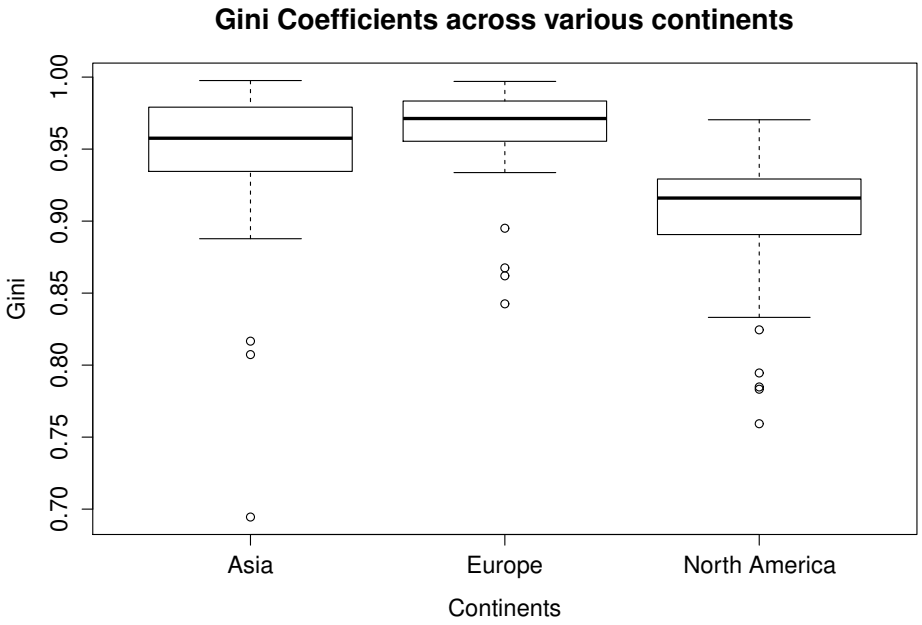


Figure 5.7: Box plots of continental Gini coefficient values

5.2.1 Gini Coefficients Across Measures

These tests investigate the Gini coefficients across continents and City Structures.

Gini Coefficients Across Continents

The box plot in Fig. 5.7 shows the distribution of the Gini coefficients by continents; as explained earlier, for North American cities have Gini coefficients that are lower than those of Asian and European cities.

The median, upper quartile and lower quartile values for North American cities are lower than those of Asian and European cities. Although the whiskers of all plots overlap, North American cities generally appear to have lower values.

A Kruskal-Wallis test of the various groups revealed that there was a sig-

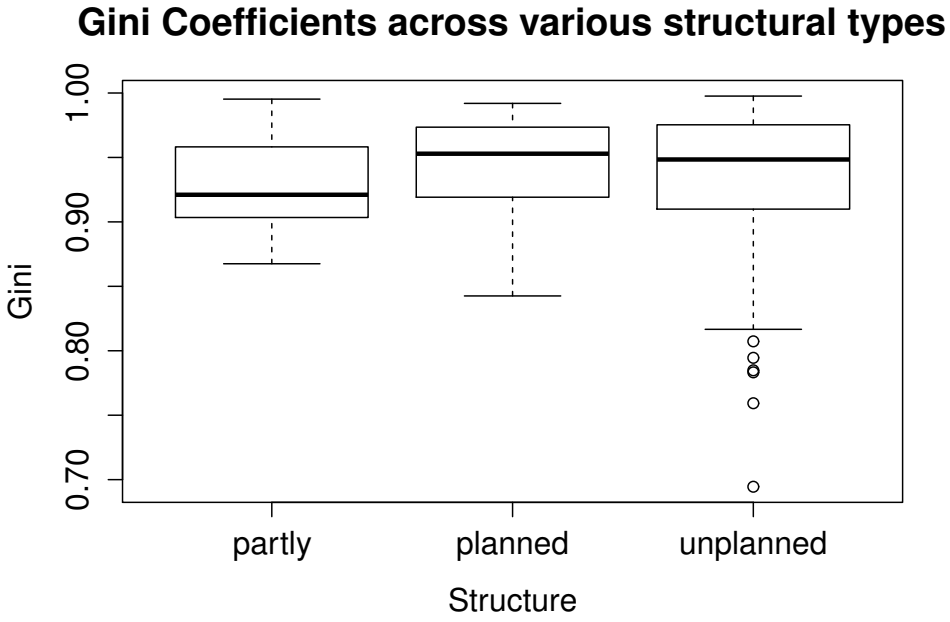


Figure 5.8: Box plots of Gini Coefficient values for City Structural Types

nificant difference between the Gini coefficient values ($H(2) = 44.6593, p < 0.05$). There is a significant statistical difference between the Gini distributions for North American ($N = 45, M = 0.901, SD = 0.48$) and Asian ($N = 28, M = 0.939, SD = 0.42$) cities as well as North American and European cities ($N = 47, M = 0.962, SD = 0.49$); however, there is little or no significant difference between the values for Asian and European cities.

Gini Coefficients Across City Structures

The Gini coefficients do not vary much across the various structural types; the box plots 5.8 make this obvious as the boxes for all three categories overlap.

The results of the Kruskal-Wallis test of this data revealed no significant difference between the Gini coefficient values ($H(2) = 2.0718, p > 0.05$).

This is probably due to the high variance in the distribution samples sizes (planned($N = 16, M = 0.944, SD = 0.34$), partly planned($N = 15, M = 0.928, SD = 0.33$) and unplanned($N = 89, M = 0.933, SD = 0.44$)).

5.2.2 Entropy Values Across Measures

These tests investigate the entropy values across continents and City Structures.

Entropy Across Continents

We tried to prove that the entropy of North American cities is higher than those of the other two categories by showing that there is a significant difference between the three distributions.

The box plot 5.9 shows the distribution of entropy values by continents; the values for North American cities are higher than those of Asia and Europe - the only over laps occur in the whisker regions. Also, the box plots for Europe and Asia overlap significantly and have similar medians and lower quartiles.

The Kruskal-Wallis test of the groups ($H(2) = 51.6169, p < 0.05$) showed that the entropy values of North American cities ($N = 45, M = 0.901, SD = 0.49$) was significantly different from those of Asian ($N = 28, M = 0.658, SD = 0.42$) and European cities ($N = 47, M = 0.639, SD = 0.49$). Similar to the Gini coefficient results, there is no statistically significant difference between values for European and Asian cities.

Entropy Across City Structures

The entropy values do not vary a lot across structural forms; this is expected as the Gini coefficients did not display significant deviations too. The box

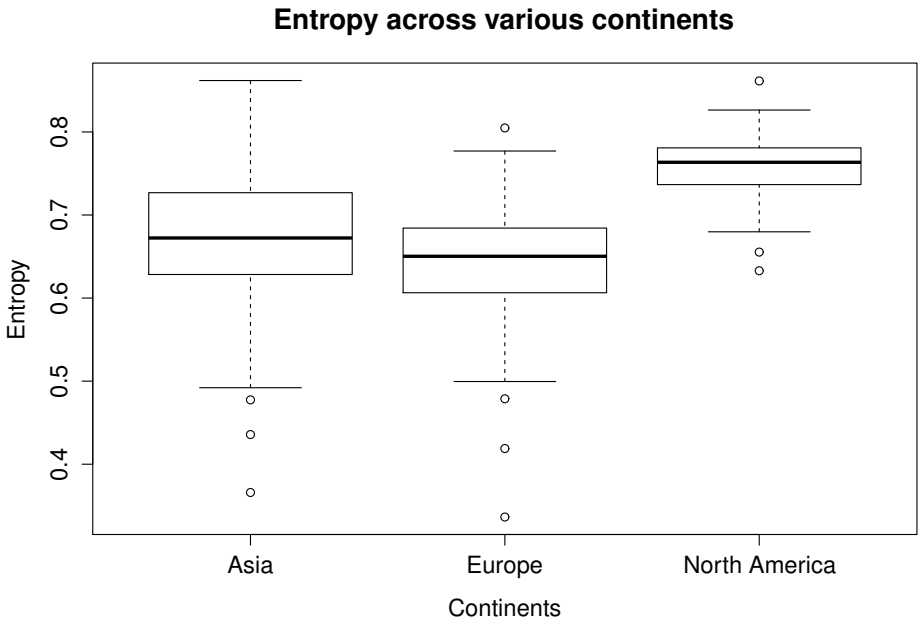


Figure 5.9: Box plots of Continental Entropy values

plots 5.10 all appear to have broadly similar upper and lower quartiles although the medians vary. It is interesting to see that unplanned cities have the highest range and also a set of outliers of extremely low entropy values; these outliers (which are all lower than the single outlier for partly-planned cities) imply that unplanned cities ($N = 89$)can be really difficult to search.

To prove that there is no major variation in the data, we run statistical analysis of the data. The Kruskal-Wallis test revealed no significant difference between the entropy values ($H(2) = 2.1987, p > 0.05$). This confirms our earlier views and can be explained by the high variance in the distribution samples sizes (planned($N = 16, M = 0.698, SD = 0.34$), partly planned($N = 15, M = 0.719, SD = 0.33$) and unplanned($N = 89, M = 0.682, SD = 0.44$)).

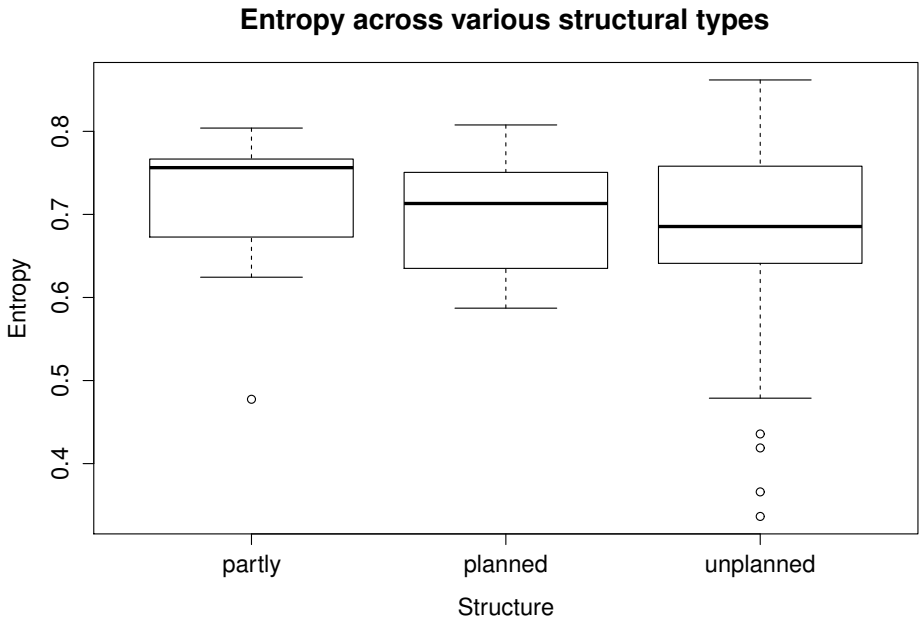


Figure 5.10: Box plots of Entropy values for City Structural Types

CHAPTER 6

Conclusion

Analysing city networks is important because it provides us with a way to measure the development of cities, identify potential hotspots, monitor the distribution and allocation of resources such as wealth, security forces and influences the planning and development of new infrastructure. It also provides us with a way to identify the challenges citizens might face in terms of mobility and transportation and enables models that aim to solve this to be built.

This research presents the results of the complexity analysis of various cities around the world. Cities have varying layouts and designs - some are highly structured while others can be random; finding out the complexities of the city layouts is important and very useful in fighting crime, city planning and in search and rescue missions.

The complexity of a city was calculated by analyzing the dual graph of the city's road layout retrieved from openstreetmaps. The dual graph ap-

proach was chosen for complexity analysis because it enhances the detection of latent information while still preserving topographical characteristics.

The eigenvector centrality for each graph serves as a measure of the relative importance of each road. The entropy of the normalized eigenvector centrality distribution of a city provided a metric for comparison. Results showed that newer cities have lower entropy values in comparison to older cities. Also, we found out that property crime is correlated to the dispersion of eigenvector centralities across the mean; crime-infested areas containing streets which have roughly the same importance ratings tend to have disproportionately high levels of crime.

6.1 Applications

The complexity analysis described in this thesis will be extremely useful to planners as it enables them see the potential outcome of new city designs and how planning decisions will affect the inhabitants, these models can also help in planning 'perfect' cities and in the early identification of potential trouble spots and areas. Such information can come in handy in search and rescue missions and in finding fugitives.

Businesses can take advantage of these models to determine the best locations to place billboards and adverts. Some businesses might benefit from adverts placed in highly-central locations while others might be more suited to areas that are remote.

This approach can be put to excellent use in the disaster response teams and aid organizations to predict the complexity of a city by using another city as a baseline; as such, it becomes easier to make estimates and plan relief distribution and allocation. Our methods can be extended to other fields (e.g. biological networks) which require the evaluation of networks

across a wide variety of criteria.

APPENDIX A

Extra Plots

This appendix includes the plots that were generated during the process runs. The figures below show scatter plots of the entropy and Gini coefficient distributions for the cities in the study.

The cities of the tag challenge have been annotated and cities are coloured by their continents.

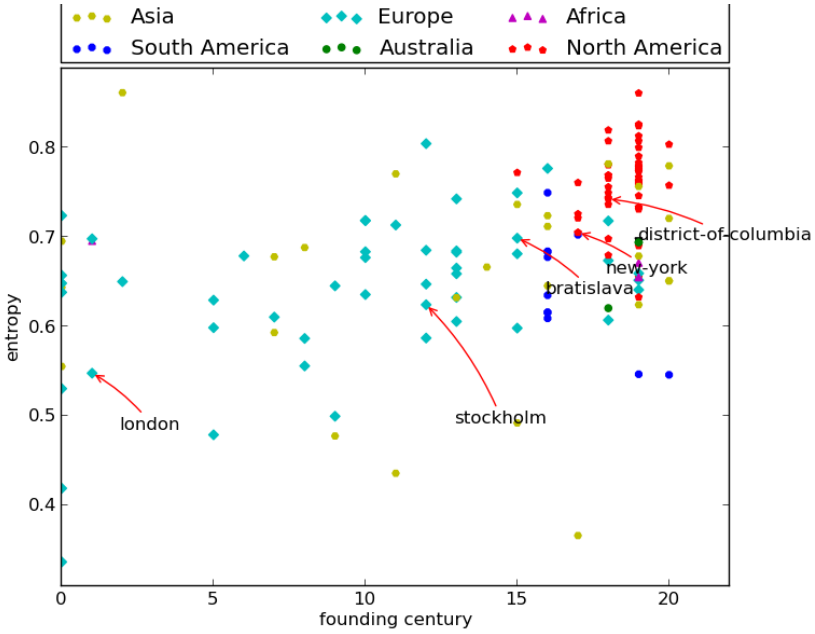


Figure A.1: Entropy vs founding century for 136 cities

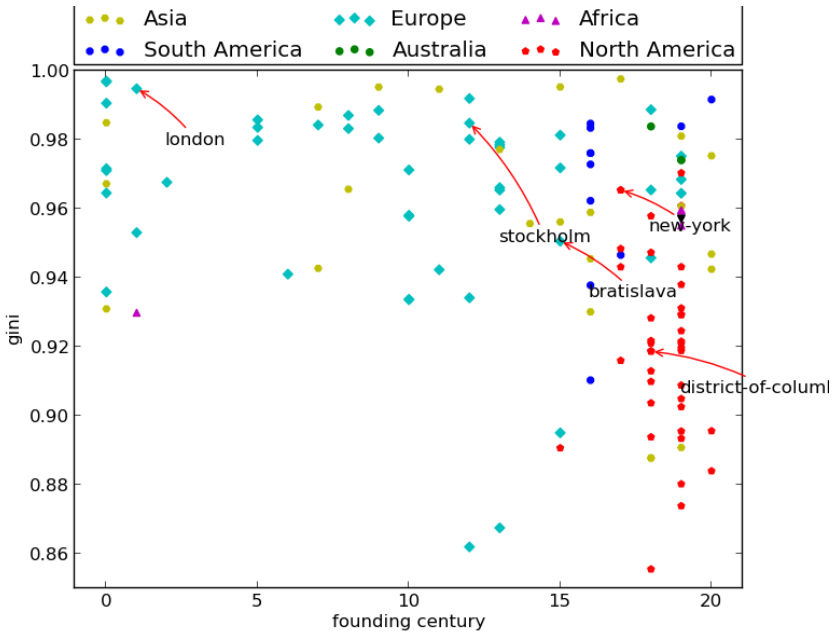


Figure A.2: Gini vs founding century for 136 cities

APPENDIX B

Abbreviations

CSV Comma Seperated Values

ICN Intersection Continuity Negotiation

ICT Information Communication Technology

Bibliography

[1] Demographia world urban areas (world agglomerations), 2012.

[2] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[3] M. Barthélemy and A. Flammini. Modeling urban street patterns. *Physical Review Letters*, 100(13):138702, 2008.

[4] M. Batty. The size, scale, and shape of cities. *Science*, 319(5864): 769–771, 2008.

[5] A. Bavelas. A mathematical model for group structures. *Human Organization*, 7(3):16–30, 1948.

[6] P. Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, pages 1170–1182, 1987.

[7] J. Buhl, J. Gautrais, N. Reeves, R. V. Solé, S. Valverde, P. Kuntz, and G. Theraulaz. Topological patterns in street networks of self-organized

urban settlements. *The European Physical Journal B-Condensed Matter and Complex Systems*, 49(4):513–522, 2006.

[8] P. Crucitti, V. Latora, and S. Porta. Centrality in networks of urban streets. *Chaos: an Interdisciplinary Journal of Nonlinear Science*, 16 (1), 2006.

[9] P. Crucitti, V. Latora, and S. Porta. Centrality measures in spatial networks of urban streets. *Physical Review E*, 73(3):036125, 2006.

[10] E. Depoortere and V. Brown. Rapid health assessment of refugee or displaced populations, 2006.

[11] P. Earle. Earthquake twitter. *Nature Geoscience*, 3(4):221–222, 2010.

[12] Ö. Esmer. *Information theory, entropy and urban spatial structure.* PhD thesis, Middle East Technical University, 2005.

[13] B. J. Fogg. Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI'03 extended abstracts on Human factors in computing systems*, pages 722–723. ACM, 2003.

[14] L. C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[15] R. Goolsby. Social media as crisis platform: The future of community maps/crisis maps. *ACM Transactions on Intelligent Systems and Technology*, 1(1):7, 2010.

[16] D. Guha-Sapir, F. Vos, and R. Below. Annual disaster statistical review 2011.

[17] D. Guha-Sapir, F. Vos, R. Below, and S. Ponserre. Annual disaster statistical review 2010. *Centre for Research on the Epidemiology of Disasters*, 2011.

[18] L. T. Gunawan, S. Fitrianie, W. Brinkman, and M. Neerincx. Utilizing the potential of the affected population and prevalent mobile technology during disaster response: Propositions from the literature. *Proc. of ISCRAM*, 2012.

[19] A. Gupta and P. Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, page 2. ACM, 2012.

[20] R. Hamaina, T. Leduc, and G. Moreau. A structural analysis of the streets network to urban fabric characterization.

[21] J. Heinzelman and C. Waters. Crowdsourcing crisis information in disaster-affected haiti. 2010.

[22] V. Hester, A. Shaw, and L. Biewald. Scalable crisis relief: Crowdsourced sms translation and categorization with mission 4636. In *Proceedings of the first ACM symposium on computing for development*, page 15. ACM, 2010.

[23] B. Hillier and J. Hanson. *The social logic of space*, volume 1. 1984.

[24] W. R. G. Hillier. From research to design: re-engineering the space of trafalgar square. *Urban Design Quarterly*, (68):35–37, 1998.

[25] A. L. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3):248–260, 2009.

[26] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer. Social phishing. *Communications of the ACM*, 50(10):94–100, 2007.

[27] B. Jiang. A topological pattern of urban street networks: universality and peculiarity. *Physica A: Statistical Mechanics and its Applications*, 384(2):647–655, 2007.

[28] B. Jiang and C. Claramunt. Topological analysis of urban street networks. *Environment and Planning B*, 31(1):151–162, 2004.

[29] K. Keizer, S. Lindenberg, and L. Steg. The spreading of disorder. *Science*, 322(5908):1681–1685, 2008.

[30] V. Lampos, T. De Bie, and N. Cristianini. Flu detector-tracking epidemics on twitter. *Machine Learning and Knowledge Discovery in Databases*, pages 599–602, 2010.

[31] A. P. Masucci, D. Smith, A. Crooks, and M. Batty. Random planar graphs and the london street network. *The European Physical Journal B-Condensed Matter and Complex Systems*, 71(2):259–271, 2009.

[32] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 international conference on Management of data*, pages 1155–1158. ACM, 2010.

[33] M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics*, pages 71–79. ACM, 2010.

[34] E. Mustafaraj and P. Metaxas. From obscurity to prominence in minutes: Political speech and real-time search. 2010.

[35] M. E. J. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[36] I. Omer and B. Jiang. Topological qualities of urban streets and the image of the city: A multi-perspective approach. 2008.

[37] L. Palen and S. B. Liu. Citizen communications in crisis: anticipating a future of ict-supported public participation. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 727–736. ACM, 2007.

[38] L. Palen, S. Vieweg, J. Sutton, S. B. Liu, and A. L. Hughes. Crisis informatics: Studying crisis in a networked world.. In *Proceedings of the Third International Conference on E-Social Science*, 2007.

[39] S. V. Pemmaraju and S. S. Skiena. *Computational discrete mathematics: combinatorics and graph theory with mathematica*. Cambridge University Press, 2003.

[40] G. Pickard, W. Pan, I. Rahwan, M. Cebrian, R. Crane, A. Madan, and A. Pentland. Time-critical social mobilization. *Science*, 334(6055): 509–512, 2011.

[41] S. Porta, P. Crucitti, and V. Latora. The network analysis of urban streets: A dual approach. *Physica A: Statistical Mechanics and its Applications*, 369(2):853–866, 2006.

[42] S. Porta, P. Crucitti, and V. Latora. The network analysis of urban streets: a primal approach. *Environment and Planning B: Planning and Design*, 33(5):705–725, 2006.

[43] S. Porta, P. Crucitti, and V. Latora. Multiple centrality assessment in

parma: a network analysis of paths and open spaces. *Urban Design International*, 13(1):41–50, 2008.

[44] C. H. Procopio and S. T. Procopio. Do you know what it means to miss new orleans? internet communication, geographic community, and social capital in crisis. *Journal of Applied Communication Research*, 35 (1):67–87, 2007.

[45] E. L. Quarantelli. Disaster related social behavior: Summary of 50 years of research findings. 1999.

[46] I. Rahwan, S. Dsouza, A. Rutherford, V. Naroditskiy, J. McInerney, M. Venanzi, N. Jennings, and M. Cebrian. Global manhunt pushes the limits of social mobilization. *Computer*, 2012.

[47] A. Rapoport. Spread of information through a population with socio-structural bias: Iii. suggested experimental procedures. *Bulletin of Mathematical Biology*, 16(1):75–81, 1954.

[48] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Detecting and tracking the spread of astroturf memes in microblog streams. *ArXiv preprint arXiv:1011.3768*, 2010.

[49] M. Rosvall, A. Trusina, P. Minnhagen, and K. Sneppen. Networks and cities: An information perspective. *Physical Review Letters*, 94(2): 28701, 2005.

[50] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[51] S. Scellato, A. Cardillo, V. Latora, and S. Porta. The backbone of a city. *The European Physical Journal B-Condensed Matter and Complex Systems*, 50(1):221–225, 2006.

[52] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[53] B. R. Sood, G. Stockdale, and E. M. Rogers. How the news media operate in natural disasters. *Journal of Communication*, 37(3):27–41, 1987.

[54] E. Strano, M. Viana, A. Cardillo, L. Costa, S. Porta, and V. Latora. Urban street networks: a comparative analysis of ten european cities. *ArXiv preprint arXiv:1211.0259*, 2012.

[55] J. Sutton, L. Palen, and I. Shklovski. Backchannels on the front lines: Emergent uses of social media in the 2007 southern california wildfires. In *Proceedings of the 5th International ISCRAM Conference*, pages 624–632, 2008.

[56] A. H. Tapia, K. Bajpai, B. J. Jansen, J. Yen, and L. Giles. Seeking the trustworthy tweet: Can microblogged data fit the information needs of disaster response and humanitarian relief organizations. In *Proceedings of the 8th International ISCRAM Conference*, pages 1–10, 2011.

[57] C. Torrey, M. Burke, M. Lee, A. Dey, S. Fussell, and S. Kiesler. Connected giving: Ordinary people coordinating disaster relief on the internet. In *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, pages 179a–179a. IEEE, 2007.

[58] A. Turner, A. Penn, and B. Hillier. An algorithmic definition of the axial map. *Environment and Planning B: Planning and Design*, 32(3): 425–444, 2005.

[59] S. Vieweg. Microblogged contributions to the emergency arena: Discovery, interpretation and implications. *Computer Supported Collaborative Work*, pages 515–516, 2010.

[60] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1079–1088. ACM, 2010.

[61] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[62] E. W. Weisstein. Grid graph. http://mathworld.wolfram.com/GridGraph.html, 2013. [Online; accessed 04-April-2013].