Ambiguity resolution of place names:
Detecting event names associated with place and phonetic conflicts in street names

CHAN Chun Keith

SUBMITTED IN TOTAL FULFILMENT OF THE REQUIREMENTS OF THE DEGREE OF MASTER OF GEOMATIC ENGINEERING (RESEARCH)

Department of Geomatics
The University of Melbourne
Australia

Feb 2014

Supervisor:
Prof Stephan Winter
Dr Maria Vasardani

(Produced on archival quality paper)
ABSTRACT

Natural language is flexible and dynamic, creating uncertainty when communicating location information. This research looks at placename ambiguity from two perspectives; (1) regarding detecting event names associated with place and, (2) regarding street name phonetic similarity. Topic one looks into how names of events and activities can potentially supersede an official place name in representing a location. Current spatial databases struggle in reconciling unofficial labels with a known location. This work explores the concept of place-event substitution and methods to detect them via social media, with the long-term goal of developing a dynamic place name database. Topic two looks into how similar sounding street names can cause confusion amongst users from a phonetic and spatial perspective. Authorities frequently use the Soundex algorithm to discover potentially conflicting street names based on pronunciation. However, there are well-documented flaws in the algorithm’s poor performance in precision and recall. An algorithm is developed based on a combination of phonetic methods, string matching and street suffix evaluation to qualitatively measure street name confusion. We apply this algorithm to the Greater Melbourne street network to uncover potential spatial and phonetic trends of street names. The expected outcome is a better understanding of the characteristics, distribution and similarity levels of street name confusion in a street network. This could further aid current location based services (LBS) that rely on street names to disambiguate location and providing useful guides for authorities in future street naming efforts.

The content of this thesis has been presented in two journal articles


DECLARATION

This is to certify that

i. The thesis comprises only my original work towards the Masters except where indicated in the preface,

ii. Due acknowledgment has been made in the text to all other material used,

iii. The thesis is less than 50 000 words in length, exclusive of tables, maps, bibliographies and appendices.

February 5th, 2014

Chun Keith CHAN
ACKNOWLEDGEMENTS

The author wishes to thank Prof. Stephan Winter and Dr. Maria Vasardani for their support and guidance throughout the entire research project, and also Dr. Martin Tomko for his continual support as the chair of the progress meetings.
CHAPTER 1

INTRODUCTION .................................................................................................................. 1

CHAPTER 2

TOPIC 1: LEVERAGING TWITTER TO DETECT EVENT NAMES ASSOCIATED WITH PLACE

2.1 Literature review on place name usage and collection .................................................. 4
   2.1.1 Traditional gazetteers .......................................................................................... 6
   2.1.2 Geospatial systems .............................................................................................. 7
   2.1.3 Non-topographical datasets .................................................................................. 8
   2.1.4 Discovering location and place names via information retrieval methods ............... 8
   2.1.5 Social media as a source of data ............................................................................. 10
   2.1.6 Social media in detecting spatial-temporal information ......................................... 11

2.2 Understanding place names and event-driven substitution ........................................... 12
   2.2.1 Scope of EDS ......................................................................................................... 13
   2.2.2 Twitter as a data source ....................................................................................... 14
   2.2.3 EDS detection & validation .................................................................................... 15
   2.2.4 Discovering unknown event names ....................................................................... 15

2.3 Detecting and modelling place name change ................................................................. 16
   2.3.1 Data collection and preparation ............................................................................. 17
   2.3.2 Experiment set up .................................................................................................. 17
   2.3.3 Burst detection ...................................................................................................... 18
   2.3.4 Event name discovery ............................................................................................ 20
   2.3.5 Substitution verification & monitoring ................................................................... 21

2.4 Experiment results ........................................................................................................ 21
   2.4.1 Burst detection via data streaming ......................................................................... 22
   2.4.2 Overall burst detection results ................................................................................ 23
2.4.3 Performance of event name discovery .......................................................... 25

2.5 Discussion of detection results, areas of refinement and limitations...................... 27

  2.5.1 Overall event detection performance ..................................................................... 27
  2.5.2 Analysis of parameter modelling techniques .......................................................... 28
  2.5.3 Challenges in natural language and Twitter talk ..................................................... 30
  2.5.4 Social modelling – accuracy & implications .......................................................... 32

CHAPTER 3

TOPIC 2: GETTING LOST IN CITIES: SPATIAL PATTERNS OF PHONETICALLY
CONFUSING STREET NAMES

3.1 Literature review on street name similarity and efforts to mitigate them .................. 36

  3.1.1 The role of street names in an organized urban environment ................................. 36
  3.1.2 Implications of street name confusion .................................................................. 37
  3.1.3 Action plans and challenges to implementing solutions .......................................... 38
  3.1.4 Soundex & phonetic similarity ............................................................................ 39
  3.1.5 Limitations of Soundex ......................................................................................... 40
  3.1.6 Alternative phonetic algorithms .......................................................................... 41
  3.1.7 String similarity .................................................................................................. 43
  3.1.8 Fusion methods .................................................................................................. 44
  3.1.9 Differences of street names vs. generic names ...................................................... 40

3.2 Designing a street name similarity algorithm ............................................................ 46

  3.2.1 Components of street name similarity ................................................................. 46

    3.2.1.1 Algorithm component (1): Metaphone ......................................................... 47
    3.2.1.2 Algorithm component (2): EndCode ............................................................ 48
    3.2.1.3 Algorithm component (3): String matching .................................................. 51
    3.2.1.4 Algorithm component (4): Street suffix ...................................................... 52
    3.2.1.5 Final scoring .................................................................................................. 53

  3.2.2 Phonetic and spatial trends of conflicting streets .................................................... 54

3.3 Algorithm evaluation setup ....................................................................................... 55

  3.3.1 Experiment design: Human-annotated test .......................................................... 55

    3.3.1.1 Dealing with street suffixes .......................................................................... 58
3.3.2 Application of algorithm on Greater Melbourne street name network .................. 59
  3.3.2.1 Handling street name particulars & exact duplicates ................................. 59
  3.3.2.2 Conflict score acceptance threshold ...................................................... 60
  3.3.2.3 Processing implementation ...................................................................... 61

3.4 Experiment results ............................................................................................ 61
  3.4.1 Performance results of algorithms ............................................................... 61
  3.4.2 Ranking results ............................................................................................ 63
  3.4.3 Overall Greater Melbourne street network conflict ........................................ 64
  3.4.4 Distribution of matches per score threshold (graph) ...................................... 66
  3.4.5 Conflicts by region (moving window results) ................................................. 66
  3.4.6 Phonetic trends ............................................................................................ 68
  3.4.7 Secondary duplicate clustering effect ......................................................... 69

3.5 Discussion of algorithm performance and spatial trends .................................... 70
  3.5.1 Overview of algorithm performance ............................................................ 70
    3.5.1.1 Analysis of results .................................................................................. 72
    3.5.1.2 Experiment integrity ............................................................................. 74
  3.5.2 Overview of Greater Melbourne street name network .................................... 75
    3.5.2.1 Spatial distribution of conflicts ............................................................. 76
    3.5.2.2 Phonetic patterns of conflicts ............................................................... 77
  3.5.3 Further algorithm evaluation (vowels and street category) ............................ 78

CHAPTER 4

CONCLUSIONS

4.1 Overall summary of research topics .................................................................... 80
4.2 Contributions and challenges towards more adaptive spatial databases .............. 82
4.3 Future work ....................................................................................................... 85

BIBLIOGRAPHY ....................................................................................................... 88
LIST OF TABLES

Table 2-1. Data collection.................................................................................................................. 19
Table 2-2. Flemington Racetrack detection results............................................................................. 24
Table 2-3. KFC Yum! Center detection results....................................................................................... 25
Table 2-4. Melbourne Race Carnival event name discovery ............................................................. 26
Table 2-5. KFC Yum! Center event name discovery.............................................................................. 26
Table 2-6. Examples of context filters from manual annotation used for machine learning.... 27
Table 3-1: Soundex code groupings based on articulation type ....................................................... 40
Table 3-2. Summary of major flaws of Soundex.................................................................................. 41
Table 3-3. Summary of prominent alternative phonetic algorithms.................................................. 43
Table 3-4. Examples of variations in vowel sounds......................................................................... 49
Table 3-5. EndCode grouping ............................................................................................................ 50
Table 3-6. Example of similarity scoring for two streets using the new algorithm .................... 53
Table 3-7. Example of step count for reorganizing days of the week .............................................. 57
Table 3-8. Street name characteristics and plan of action................................................................. 60
Table 3-9. Comparison of precision and recall values for all algorithms....................................... 62
Table 3-10. Similarity results for “Craig Street” as produced by Mturk and all algorithms........... 63
Table 3-11. Ranking performance of Levenshtein Distance and new algorithm.......................... 64
Table 3-12. Sample ranking for Levenshtein Distance and the new algorithm............................... 59
Table 3-13: The top occurring phonetics and associated street names.......................................... 68
Table 3-14. Trends based on the comparison of the original street name with its associated matches ........................................................................................................................................... 72
Table 3-15. Ranking of street suffix based on user perception........................................................ 73
LIST OF FIGURES

Figure 2-1. Twitter stream of Crown Oaks Day verification ......................................................... 22
Figure 3-1. Example of bigrams (n=2) which creates 5 groups of letters ...................................... 43
Figure 3-2. The 16 consonant sound categories of Metaphone .................................................. 47
Figure 3-3. IPA diagram for vowel sounds ..................................................................................... 49
Figure 3-4: Example of moving window concept for street similarity assessment in a region ..... 54
Figure 3-5: A sample of Mturk survey questions for Diana Street .............................................. 56
Figure 3-6: Overall precision/recall results based on average of all four sets ......................... 62
Figure 3-7. Visualization of full street name conflict levels of Greater Melbourne ..................... 65
Figure 3-8. Street name conflict zoomed-in to Melbourne CBD ................................................. 65
Figure 3-9. Distribution of matches for thresholds of 0.5, 0.7 and 0.9 ........................................... 66
Figure 3-10. Moving window results ............................................................................................. 67
Figure 3-11. Aggregated results based on municipality with street density influence normalized 67
Figure 3-12. Conflict count based on phonetic code length ....................................................... 69
Figure 3-13. Analysis of secondary duplicate clustering on results ............................................. 70
CHAPTER 1
INTRODUCTION

People have developed labelling systems for urban environments in order to locate places and navigate. These labels act as a form of identification and are designated to buildings, streets, blocks, neighbourhoods, landmarks and any form of territory considered significant enough to index. They provide a practical alternative to the more precise but less relatable coordinate system (e.g., satellite positioning). Thus, as opposed to memorising a latitude/longitude pair, people communicate spatial information by referencing names of known places.

However, communication of these labels in natural language can be inaccurate, imprecise and flexible. Inaccuracy can be caused by mistakes during the process of hearing a label or retaining it to memory, leading to an imperfect knowledge of a label (dropped, added or altered particles). Imprecision occurs when there is accurate knowledge of the label but it is poorly reproduced in oral communication due to physical (speech impediments, poor articulation) or cultural (dialect, slang, language) influences. Flexibility is a result of intentional choice to use an alternative label above an existing official one. Thus, there can be a myriad of situations that can lead to a breakdown in understanding when communicating location. This research looks at two aspects of resolving problems in location communication in an urban environment, both with their unique causes of conflict:

**Topic 1: Leveraging twitter to detect event names associated with place**

Place names are a type of label assigned to identify geographical regions, of which they can be official (bestowed and recognized by authorities/custodians) or unofficial (created and propagated by on-the-ground users). Place name databases function as repositories for these labels and primarily store official place names. However, it is common for a society to stray from this rigid labelling system and develop alternatives. Such alternative place names can have a short lifespan, of which official gazetteers fail to detect on a regular and timely basis. Thus, the current place name knowledge is incomplete and generally less adaptable to these terminology variations. This topic looks into the potential of leveraging social media to detect and uncover these alternatives in place name terminology. This could lead to a more authentic
perspective of place names usage that can be obtained and monitored in a close to real-time manner.

**Topic 2: Getting lost in cities - spatial patterns of phonetically confusing street names**

Street names are an example of a spatial indexing system designated and managed by authorities as a component of the urban address system. However, it is common to find many duplicate and similar sounding street names in urban street networks. Confusion can arise when two parties are attempting to verbally resolve location ambiguity using street names with the prospect of close sounding alternatives in their vicinity. Information transfer is hampered due to the probability of mistaking one street for another or an overall difficulty in distinguishing between a collection of similar options. This topic explores street name confusion from the perspective of phonetics and spatial distribution by utilizing a street name similarity algorithm. The analysis from this topic can lead to better methods in evaluating and mitigating street name confusion in a street network.

Hence, the overall research question of this thesis is:

- Can we develop systems/solutions to augment our spatial databases in order to improve on ambiguity resolution of place names communicated by people?

In both communication aspects, there is a need to resolve ambiguity of correspondence in order to determine location; therefore, they share similar motivations and repercussions. Ambiguity resolution in such cases will be an additional step towards bridging the divide between human-machine interactions. Since most information and records are now digitally stored, there is a need for the machines to effectively traverse between the data source (e.g., databases, servers, digital libraries, etc.) and the user (human). A major challenge to this end is the ability to easily adapt to user input. This includes the ability to resolve uncertainty in terminology usage and speech. Subsequently, this research would aid in services reliant on location, especially emergency response teams. These services rely heavily on verbal cues to properly dispatch and locate within an urban environment. Failure to disambiguate place names can cause loss in time, money, productivity and even lives (Coveyduck et al. 2004; Ciccone 2012; Flowers 2012). Additionally, a more in-depth understanding of the problems associated with labelling would provide authorities and policy makers a more effective guide to develop and manage the labelling methods within an urban environment. Hence, the
overall long-term goal is for advancements in spatial datasets that can easily handle vagueness and respond with "did you mean to say...?" or “is your location also known as...?” based on appropriate user input and context.

CHAPTER 2

TOPIC 1: LEVERAGING TWITTER TO DETECT EVENT NAMES ASSOCIATED WITH PLACE

In human communication it is common to stray from a rigid structure of labelling when referring to a place, based on a sense of informality. Regardless of whether an official place name exists and/or is known, people may intentionally use alternative terminology. The label usage itself cannot be enforced in casual day-to-day conversations and is dependent on personal choice. Thus, the value of a place name is derived from the level of recognition and usage it holds within a community.

A population collectively knows these changes, but they can remain hidden from current spatial databases that are more static in nature and based on terminology from assigned contributors, or restricted maintenance policies. Over time, these databases inevitably fall short in relevance and breadth of vernacular. However, with the proliferation of social media usage, particularly Twitter, there is an opportunity to directly observe and assess how the public is choosing to label the places they know. Hence, the hypothesis for this topic is that social media can be leveraged to create a dynamic place name database that can detect and recognize when place names are being substituted by an alternative name in public communication.

For this topic, the place names in question encompass places of interest (POI) such as businesses, buildings, activity centres and man-made constructions, all of which possess a specific name or identity. In terms of place name substitution, this topic looks particularly at scenarios triggered by human-centric events. Multi-purpose establishments, like amphitheatres and arenas, are unique because while they carry official names, their function
as a host to continually changing clients can strongly influence their identity. These client events can overpower the official place name in everyday communication of locations, and expressions such as “in stadium X” can be substituted with “at concert Y” as a locative expression.

2.1 Literature review on place name usage and collection

Labels are used to describe/identify topological features, points of interest, or demarcated geographical regions in our environment. Most places and locations hold official labels, called place names. These place names are assigned and administered by an authoritative body, usually a government agency, tasked to validate and authenticate these labels (Hill 2000). The place names themselves can have many origins, including citizen feedback, historical significance, or descriptive in nature. These place names are expected to be known by the general public and used for all official purposes. However, on a day-to-day basis and within informal settings, it is common to find people straying from these official place names, even when they are known.

This phenomenon can be linked to the tendency for a society to naturalize their environment via language. For example, when interacting with someone for the first time, people tend to be more formal and complete with names and titles in recognition of social hierarchy and mutual respect (Brown et al. 1968; Wood and Kroger 1991). However, over time, familiarity causes labels to evolve, often dropping official titles first, moving on to single name basis, and sometimes even to nicknames for close friends or family (Van Buren 1974). This is connected to the change of status, power and structure between the two subjects involved. Thus, the closer a relationship is, the more room there is for informal naming or nicknames (Leslie and Skipper Jr 1990). Each subsequent label evolution represents a change in the relationship level with the subject, which can be seen as a result of human nature rather than necessity.

The evolving of terminology in human communication also extends to place names, with a tendency to de-formalize labels as we become more connected with the place. A place itself is just a defined spatial region and might seem un-relatable to humans, however the activity it hosts can create an emotional attachment (Kostanski 2011), especially if it occurs within a social context. Zhou et al. (2005) discusses the concept of describing the same place differently
depending on the audience and the intention of the user, which in turn can be influenced by various factors such as relationship levels, shared understanding and knowledge or area. Thus, since place comes into being when space is assigned meaning (Carter and Donald 1993), people are not just travelling between points and coordinates, but rather regions of personal interest that connect us to each other, with each label variation the product of intention and purpose. The identities of these places are shared among the general public, whereby over time certain terminology will prevail in usage.

The many types of alternative place names are based on the activity involved and the characteristics of the community utilizing it. Venues that act as a home ground for sports teams can acquire nicknames created and propagated by their fans. For example, “Theatre of Dreams” (Manchester United’s Old Trafford), “The Grove” (Arsenal’s Emirates Stadium) and “Gallowgate” (Newcastle United’s St. James Park) are nicknames of English Premier League stadiums that were coined and eventually adopted by their respective followers. This concept of topophilia (strong attachment to place) is common for sporting grounds which are often considered sacred to fans (Antonsich 2010). This strong connection between community, place, and place name has even lead to public dissent and open demonstrations when commercial sponsors attempt to gain naming rights of these sports venues (Boyd 2000; Duke 2002; Edwards 2012). This tension is created by the desire to uphold tradition and history of a club versus the potential for the club to grow via the amalgamation of commercial entities into their identity via sponsorship deals (Vuolteenaho and Kolamo 2012). Thus, conflict of identity occurs, resulting in the threat of authentic culture (football native landscape) being taken away along with the original labels and place names used (Vuolteenaho and Kolamo 2012). As a whole, this demonstrates the importance of place names as more than just a generic label of place, and its significance to community identity.

Nicknames can also be casual and generic when used within the context of a smaller, tight-knit social group participating in a regular activity. For example, a family can go to “the beach”, schoolmates would visit the local “mall”. In these scenarios, the shared knowledge and understanding of context allows for place name specifics to be dropped. Finally, alternative names can rise from the modification or splicing of original names based on cultural and linguistic influences. For example, the term “Maccas” is common slang in Australia as a reference to fast-food chain McDonalds. Interestingly enough, this Australian franchise recently chose to adopt the slang as the new official title in 2012. Granted that it was a
marketing strategy, it still highlights the relevance of recognizing crowd-generated alternative place names and their influence on a community. In all these cases, prominent terminology might not necessarily endure, and can be superseded for short or long term durations based on circumstance. Although a population collectively knows the changes, they can remain hidden from current spatial databases. Section 2.1.1 to Section 2.1.4 introduces the concept of gazetteers, general spatial databases, non-topographical datasets and latest information retrieval collection techniques along with their limitations in dealing with these temporally dynamic place names. Next, section 2.1.5 and Section 2.1.6 looks into the potential of social media as a source of geospatial data and the latest work that are related to this area.

2.1.1 Traditional gazetteers
Historically, official place names related to geographical or physical features have been collected and collated in gazetteers. Gazetteers function not just as a long-term centralized database of place names but also as points of reference to tie place names with locations, a process called geocoding (Goldberg et al. 2007). The location information is usually stored as a pair of two-dimensional coordinates and a basic entry consists of a place name, coordinate, and type (Hill 2000). Gazetteers function as an authoritative dataset and sources data from more established organizations like government agencies and geographic bodies.

For official gazetteers, various government bodies/agencies are charged with their administration and maintenance. An example would be the Gazetteer of Australia managed by the Intergovernmental Committee on Surveying and Mapping (ICSM), the contents of which can be accessed via the online portal Geoscience Australia (Intergovernmental Committee on Surveying and Mapping, 2014). The task of collating and maintaining place names is sometimes delegated to state level. The official register of Geographic Names for the state of Victoria is centralized under the VICNAMES database (Department of Sustainability & Environment, 2011). This database includes all geographic features, administrative localities/regions and roads. Worldwide, There are many prominent agencies charged with official gazetteers. The United States Board on Geographic Names (UBGN) (The U.S. Board on Geographic Names, 2014) and the Ordnance Survey (United Kingdom) (Ordnance Survey, 2014) is in charge of most spatial data for their respective countries, including place name data. For foreign regions (all area outside the borders of a country), the US has The National Geospatial-Intelligence Agency. This agency maintains a database named GEOnet Names Server (GNS), which is an official repository of foreign place names approved by UBGN (National Geospatial-Intelligence
Agency, 2014). These databases are used for, among other things, emergency services, navigation, routing and town planning. They can potentially contain information on official secondary names (dual-named places) or widely accepted alternate names, but by default do not contain place name vernacular. Thus, the breadth of content of gazetteers is limited by its function and data source. Additionally, administrative or bureaucratic procedures can often lead to slow update rates.

### 2.1.2 Geospatial systems

There are a various unofficial gazetteers of place names and map data usually maintained by non-governmental organizations (NGOs) and commercial entities. The data source of unofficial gazetteers can be varied, including official gazetteers themselves (where data is publically available), mapping companies and citizens (volunteered geographic information). Some examples of the more prominent unofficial gazetteers are Geonames and Getty Thesaurus of Geographic Names (Hill 2000). More user-focused geospatial systems like GoogleMaps, Apple Maps and Bing Maps (Microsoft) combine data from different sources (official place names, points of interests and alternative labels) in an online interactive map interface. They have a higher usage rate because they are designed to be a more user-intuitive application. Data is sourced from various spatial data companies and integrated as layers and metadata into their system. The market for user-centric mapping services is considered highly competitive (Miller 2013), which drives innovation and the width of dataset variety.

GoogleMaps, as a product of the larger Google Corporation, is able to supplement its spatial search engine with proprietary data mining techniques (Carr 2006). If a search request does not match any place name in their database, the GoogleMaps algorithm will crawl through webspace to find any occurrence of terminology from online articles or reviews in an attempt to tie in a location. However, if the requested terminology is too new with a limited internet footprint, this method will also fail. Business can also manually apply to be listed in the database, but with limited participation success.

There are also spatial datasets generated from crowd-sourced methods. OpenStreetMap is a prominent community-driven spatial database with a vast corpus of data including transport networks, administrative boundaries, and business listings (Haklay and Weber 2008). Usage is open access and data is sourced from any person/party with a valid server account and the willingness to contribute. Contributors consist mostly of individuals (personally collected GPS
tracks), government agencies (public domain data) and spatial data companies (free license data) (Wikipedia, 2014a). The general quality and integrity of information is overseen by a group of senior members (super administrators). This setup is less hindered by bureaucratic procedures, enabling data updates to be implemented more efficiently. However, any changes or corrections to the dataset have to be actively observed, collected and uploaded by a willing registered user, making it unsystematic and highly dependent on participation levels (Haklay and Weber 2008). This usually confines contributions to the minority of enthusiasts (less than 30% of its two million users have made any contribution). Additionally, there have been known problems in regard to poor digitization, lack of quality control and uneven distribution of data (Haklay 2010).

Overall, the majority of these spatial databases still function as long-term repositories of places, rather than an active observer of place name changes. Event coordinators rarely see the benefit of uploading such short-term listings in these databases. They instead opt for direct marketing campaigns through advertisement on traditional media (television, radio, posters) and social media to disseminate information of event/place/time.

2.1.3 Non-topographical datasets
Non-topographical spatial datasets function as a directory of points of interests (POI) and business listings. They are a useful source of commercial entities, which is something usually ignored by traditional gazetteers. They provide a more organized and cohesive list of businesses, usually with a category-based indexed search system. They are produced in both hardcopy (e.g., yellow pages) and online versions (e.g., Yelp, Superpages and Yahoo Local) with variations of participation levels (Antin et al. 2012). Generally an entry consists of a listing name, contact and address. Online datasets are less restricted in update time, but still suffer from low participation levels because not all business are aware of such services, or interested in creating a profile. Furthermore they usually restrict their scope to fixed establishments, rather than short-term events themselves.

2.1.4 Discovering location and place names via information retrieval methods
Place name and location information can also be extracted from existing datasets in an automated/semi-automated manner using geographic information retrieval (GIR) methods. In the field of GIR, algorithms can be trained or supervised over a controlled dataset in order to make assumptions and prediction as to how to extract and interpret textual information from
any corpus of data. For example, there have been efforts to automate the entire development of a gazetteer through web crawling (Popescu et al. 2008). Here, the three components of place name, coordinates, and parent category are determined by mining publically available websites containing articles with references to place names. Naturally, this work is reliant on the level of “presence” a place name has in the searchable Internet, meaning that success is geared towards long-term existing labels, and not the possible dynamically changing alternatives. This effort, like many others, is based on the concept of geo-locating generic documents based on their textual content. This can be achieved by looking at the geographical properties of certain types of words like toponyms (e.g., London, University of Melbourne, etc.), geographic features (e.g., mountain, creek, river, etc.), culturally local features (e.g., curling, skiing, etc.) and geographically localized slang/dialects (e.g., soda/pop/soft drink) (Roller et al. 2012). Also included in this difficult task of geo-locating documents is the issue of toponym resolution, which is the task of reconciling place name to location when the place name involved has multiple occurrences (non-unique place name) (Leidner 2008).

There have been many works over the past decade in improving document geo-location especially on online/digital documents. There were attempts to determine the geographical scope of web sources (Ding et al. 2000). Closely related is work in toponym disambiguation for digital libraries with a reported high precision (>0.74) and recall (>0.89) (Smith and Crane 2001). This involved identifying, categorizing and disambiguating place names from an online corpus of information that included historic (ancient Greece and Rome) and contemporary locations (London, California and Upper Midwest). Other research used language models and utilized a grid map with the purpose of assigning documents to the appropriate geo-cell (Wing and Baldridge 2011). This work achieved a median prediction value accuracy of 11.8km for each document. More mainstream data sources have been explored for document geo-locating, including Wikipedia. TextGrounder was a geoparsing system developed to learn the relationship between words and places on Wikipedia documents (Brown et al. 2012). This same research also continued on with the same efforts on Geo-Wiki, consisting of geotagged English articles from Wikipedia (Roller et al. 2012). Overall, most of these datasets and method were still limited by the width and depth of place names included. There is an aspect of information timeliness, especially when dealing with short-term alternative place names, which could be missing from generic online documents.
2.1.5 Social media as a source of data

Considering the limitations of the current place name data collection methods, it is valid to look at other sources. In the past decade, the rapid adoption rate of various social media platforms such as Facebook, Twitter, Flickr and Instagram has created a wealth of user generated data of daily life (Elwood et al. 2012). These platforms are designed to allow society to broadcast their life activities and thoughts to a wider audience in the form of micro-updates. These updates can take the form of a video, images or written text. Each platform has varied privacy settings in terms of what information is public domain versus what is personal content. However, the possibility of leveraging this information is akin to tapping into the collective consciousness of a society.

One of the primary characteristics of this data type is the lack of intentionality required in order for a contribution to occur. It is deemed to be less of a crowd-sourcing initiative and more of a passively volunteered system. This is both an advantage and disadvantage. Collecting passive data dramatically increases the potential amount of data contributors without having to lobby for active participation. These larger datasets will generally lead to better statistical aggregation. On the other hand, analysing the data requires sifting through mass amounts of irrelevancy and necessitates developing well-suited tools and methods, hence the term “data mining” (Bollier and Firestone 2010). Another positive is that passive data is much more attainable due to its pre-existing, open source platform. For example, Twitter offers an application programming interface (API) functionality whereby content can be easily collected using basic coding instructions. This leads to data collection that is efficient, convenient and easier to scale up. The results can also be considered a more organic and truthful representation of communication (Marwick and Danah 2011) since the collection process is not “staged” or consciously known by the users, as opposed to user polls/surveys.

On a global scale, there have been major developments in harnessing big data from social media to assist in understanding the needs of communities and organizing the delivery of aid. Some significant examples are Global Pulse, a United Nation initiative attempting to model health, social and economic patterns in various regions in order to strategize policies (Lohr 2012) and Ushahidi, an open source web-based platform, which allows for customized crisis relief maps (Gao et al. 2011). Ushahidi was utilized in the recent 2010 Haiti earthquake and 2011 Japanese Tsunami in order to identify significantly affected regions and the type of assistance needed for each area. For all these cases, social media was considered a significant
gauge of discussion trends, thus only what was relevant in a community was picked-up by the system (Gao et al. 2011; Lohr 2012).

However leveraging social media data is not without some limitations. For example, when it comes to disaster management, there has been difficulty in integrating crowd-sourced information (e.g., microblogs) with the existing systems of crisis organizations (Gao et al. 2011). This is mainly due to questions of the integrity, accuracy timeliness and security of information provided (Truelove et al. 2014). Additionally, volunteered geographic information (VGI) can potentially generate skewed data due to the given sample set of users utilized. The activity levels of VGI have shown to be co-related to demographics, region, and social-economic background of the community involved (Haklay 2010; Haklay et al. 2010). Hence this form of data has to be interpreted and used with caution.

2.1.6 Social media in detecting spatial-temporal information

There have been attempts to spatially refine results of non-geocoded social media data by developing location-defining methods. Proposed solutions included analysis of metadata and time zones (Davis Jr et al. 2011; Hecht et al. 2011; Li et al. 2011), and regionalizing based on terminology/slang (Cheng et al. 2010; Gelernter and Mushegian 2011; Kinsella et al. 2011; Baldwin et al. 2012; Eisenstein et al. 2012; Gonzalez and Chen 2012). These methods can resolve location to city level (at best) and with limited precision. This is due to the low participation rate of users willing to activate and share location via the in-built location-sharing option (Davis Jr et al. 2011; Gonzalez and Chen 2012). Hence most location resolution efforts have focused on examining long-term trends and tweet content in user discussions in order to determine a generalized location (Cheng et al. 2010; Davis Jr et al. 2011; Boettcher and Lee 2012; Gonzalez and Chen 2012). More importantly, these various research topics were focused on where people are, and not how they are communicating the places they visit and the terminology involved. Additionally, named entities can be mined and collated from social media feeds relatively efficiently with the help of basic language processing tools and a Named Entity Recognition (NER) tool (Hasegawa et al. 2004; Sakaki et al. 2010; Gelernter and Mushegian 2011; Liu et al. 2011).

There has been work in detecting temporal patterns and bursts of human-centric activities. This included the attempt at detecting regular events (cinema going, holiday celebrations) and unique events (pentagon attack, hurricane) modelled from user queries in the Microsoft
Network (MSN) search engine (Vlachos et al. 2004). Other research looked at detecting game updates in the National Football League (NFL) in the United States (Zhao et al. 2011). This research successfully detected touchdowns, interceptions and field goals at the rate of minutes to seconds by observing and leveraging the ten most common game lexicons and team names. Many twitter-related trend/event detection research also share these same principles of using phrase-count and clustering levels in their systems (Rattenbury et al. 2007; Mathioudakis and Koudas 2010; Sakaki et al. 2010). These examples were built on the assumption that the social media usage was sufficiently high and widespread to represent a community and its daily life experiences.

Besides that, there has been work on detecting local tourist-orientated activities/events by clustering tweets based on creation time, location and common keywords (Boettcher and Lee 2012) and also by grouping terminology based on common themes of people, place and purpose (Watanabe et al. 2011). However, both scenarios are limited by the necessity for tweets to be geo-referenced, thus severely limiting tweet count and event coverage. For most of their experiments, less than ten tweets were used to model a single event, of which bots or 3rd party applications could have potentially generated many of those tweets.

The closest effort related to our work, in terms of automated detection, is from the possibility of extracting event and place names from Flickr tags. This was achieved using burst detection algorithms (Rattenbury et al. 2007; Purves and Hollenstein 2010), but is limited by the characteristics of its data source. It is not as responsive to real-time changes due to the time-dependant nature of capturing, tagging and uploading photographs, thus limiting it to a post-process analysis. The act of discovering new and previously unknown terms in real-time has been relatively unexplored. In summary, for all the various research developments, there is limited progress in the task of contextualizing an event that is unknown along with its uncertain terminology in a timely manner, especially within a geographical context.

### 2.2 Understanding place names and event-driven substitution

Official place names exist based on a governing authority that validates it. As discussed in Section 2.1.1, this is usually the responsibility of relevant government agencies. An alternate place name is defined as any label used to refer to a location that deviates from an existing
official place name. Alternative place names are seen as the product of language/cultural naturalization of place by a society (as discussed in Section 2.1). The extent of deviation of a place name from its original/official form can be varied. It could be derived based on its initials, an abbreviation, a nickname or a circumstance (Henderson and Cashmere 1976). The motivations can also be varied, be it vanity, economic or culturally driven (Winfield-Pfefferkorn 2005; Wright 2012). Alternative place names can also have a life span, with some labels becoming more prominent or prolific than others over time.

This research specifically looks at alternative place names caused by event-driven substitution. An event-driven substitution (EDS) is defined as an occurrence, whereby a feature with an official place name is referenced by the temporary activity it currently hosts. In particular, the reference is for multi-purpose buildings and structures that play regular host to active human-centric activities. For example, a travelling carnival X might set up their tent in public park Y for one weekend. There might be a general knowledge of park name Y, but the prominence of carnival X might cause it to supersede the parks’ name in terms of spatial referencing. Thus people may say, “lets meet at X this afternoon” instead of “there is a gathering at Y”.

In theory, the official name is always the defining descriptor of these places and each pertaining event is just a temporary activity status held by the client. However, in practice, as occurring in human communication, event label and place label become interchangeable. In some cases, the event takes over as the nominal representation of that geographical region. The official place name is never permanently substituted in records, but rather overpowered by a temporary alternate label in real-life communication.

The assumption is that during non-events periods, a host site will be referenced by its official name in regular daily discussions, at a relatively steady state. When coming closer to an event, discussions regarding the host and the event should rise. Depending on the prominence of the event and the will of the community, discussions using the event name could surpass that of the official place name. This lifespan is generally limited to the length of the event. That phenomenon would signal that a temporary event-driven substitution has occurred.
2.2.1 Scope of EDS
The potential events that undergo substitution can be categorized into two primary types, namely regular events, and one-off events. Regular events refer to activities that regularly re-occur in the same geographical location, usually following a weekly, monthly, yearly, or other periodic cycle. One-off events are unique activities without any immediate plans for re-occurrence at the same place. While place name substitution can occur for any situation, the experiments are established based on the following theories:

Regular/cyclic event characteristics
• The consistency of the event and the constant community involvement in each iteration create a lasting imprint in communal memory.
• The positive reinforcement of participating in an activity fixed to a geographical location, leads to a strong sense of communal connection between place and event.

One-off unique event characteristics
• When the popularity of a unique activity is significant enough, it can engulf the hostname in terms of geographical representation for the period it is active.

Regardless, for these experiments it is assumed that EDSs are a distinct phenomenon that alter the status quo of place name usage, meaning the official place name will return to default usage once the event has sufficiently died down.

2.2.2 Twitter as a data source
This topic uses the publicly available Twitter data as a representation of communal thoughts occurring in real-time. From among the many social media options, Twitter has been particularly favoured by researchers due to its several key advantages. Unlike mainstream news content, alternative social media (e.g., Facebook, Tumblr, Instagram) or various other blogging sites, Twitter has the most confined content parameters. The mandatory 140-character limit eases the processing step of trying to decipher large amounts of textual content. On the other hand, this character limit might also encourage users to be more semantically thrifty, which could lead to compression, alterations and abbreviations of place names. That being said, studies have revealed that the lexical quality of text in social media is mid-range and improving over time when compared to a range of information resources (e.g. NY Times, Yahoo!, Blogs, .com domain, .edu domain) (Rello and Baeza-Yates 2012). Also,
Twitter is possibly the most influenced by trends and on-the-ground hype. While other platforms focus on creating mini blogs of a users’ life akin to a diary, Twitter encourages live updates and short discussions of a users’ current activity, resulting in each new feed quickly overshadowing previous ones. This means the value of a tweet is just as much based on its age and speediness, as its content. For that reason it is speculated that real-time events can be discovered in a matter of hours to minutes, if there are sufficient users involved. Also it is argued that it can precede other traditional information sources such as news feeds (Baldwin et al. 2012; Gelernter and Wu 2012) and opinion polls (O’Connor et al. 2010), however, the extent of this is not wholly agreed (De Longueville et al. 2009). Kwak et al. (2010) finds that while general news is picked up by mainstream mass-media sources earlier, certain types of information (e.g., vehicle accidents, sporting updates) appeared in Twitter streams first. However, Twitter’s overall characteristics of being open source, having a large user base, near real-time data streaming and controlled format make it an ideal data source for processing social discussions (Nair and Narayanan 2012).

2.2.3 EDS detection and validation
In this topic, an event name is said to have substituted a place name when in communal discussions:

(1) An event name is used without mention of the official name.
(2) An event name is used in the correct context of place-time and not as a generalized mention.
(3) The amount of discussion fulfilling both (1) and (2) surpasses that of official place name usage.

The presence or absence of place name terminology in their tweets is seen as an intentional user choice and a representation of how they perceive their world spatially. Hence, each individual tweet is considered an equally weighted vote towards a certain terminology. No bias is given towards the number of followers each user has or how many users they, in turn, are following.

2.2.4 Discovering unknown event names
Within a community, the rate of a place name being used or brought up in everyday discussions is assumed to exhibit a stable profile during downtime of non-events. This stable profile is a product of any generic/random discussions that use the place name, and of which
are unmotivated by any specific event. For example, tourists might plan a tour of a famous sports stadium, an accident might occur near a theatre or a group of friends might decide to meet outside a prominent garden. All of these scenarios would potentially generate discussions with the location place name being used, event though no significant event is being held at the location. If the frequency of occurrence of these discussions is observed over a long enough period of time, it is assumed that a consistent linear trend should appear.

This stable profile is modelled via Twitter by assuming that tweet rates exhibit a normal distribution, whereby each tweet is seen as a random variable independently drawn from a common source. The normal distribution model is a commonly applied concept in natural and social sciences, including recent works in monitoring social media trends (Asur et al. 2011) and network behaviour (Lerman and Ghosh 2010). In this scenario, the normal distribution model is the average tweet rate (per hour) recorded during non-event periods.

Any deviation from the stable profile is said to be potentially triggered by an event. Given the assumed normal distribution beforehand, this spike can be recognized by statistically comparing a current discussion profile with its past stable profile. This deviation is assumed to be caused by the following increase of content type:

(1) Official place name usage
(2) Event name usage
(3) Official and event name usage

With prior knowledge of only an official place name, discussions of type (1) and (3) are obtainable once a possible event has been detected. A sample batch of discussions during the detected event can be used to inspect for context, language and frequency of terminology in order to uncover the event name itself. Once an event name is discovered, the usage rate of both official place name and event name can be actively monitored to verify if participants are choosing one terminology over another. If this were found to be the case, it would validate the theory that the event name is taking the role of place name.
2.3 Detecting and modelling place name change

The method of data collection and necessary preparation steps are discussed in Section 2.3.1. Section 2.3.2 provides a summary and justification for the place/event scenarios chosen. The processing itself consists of three steps: (1) burst detection, (2) event name discovery, and (3) verification and monitoring. These three steps are discussed in Sections 2.3.3 to 2.3.5.

2.3.1 Data collection and preparation

Tweets are collected via the Twitter streaming API (Gardenhose), which allows for 5% throttled sampling of all occurring tweets. This limitation does not affect the completeness of data if the intended stream is not of high volume (less than 10 tweets/hour). The two most common methods for the API streaming are via keyword search or coordinate search (bounding box). Coordinate search is not used due to the low count of geo-referenced tweets (<3%), attributed to the opt-in usage policy implemented by Twitter. Additionally, the majority of these geo-referenced tweets are motivated by 3rd party apps that produce semi-automated responses (e.g., through Foursquare, Hootsuite, etc), and are thus less representative of a users’ real voice. The more favored keyword search method allows for all real-time tweets to be monitored for a specific group of word(s). Only tweets in English language were utilized. Data is collected and post-processed, but there is direct potential to implement a fully automated, real-time system.

Retweets (RT) are a non-original content duplicated and forwarded from one to many users. Since they might not accurately reflect the style and syntax of the user who forward it, they were removed from the dataset. RTs are identified based on tweet metadata. Tweets from original users are still preserved and processed. Twitter also contains high amounts of spam generated by spam bots. Spam bots are non-human Twitter accounts programmed to flood the network with content in order to aggressively push a business, product or idea, which inadvertently skews potential tweet analysis (Benevenuto et al. 2010). They gain traction by mimicking casual human conversation to deceive users and by incorporating trending terminology in its tweets to increase viewership. Spam bots innately have a high follower count and an unusually low following-to-followers ratio (Calzolari 2008; Benevenuto et al. 2010; Lee et al. 2011) due to their need to attract the biggest possible audience. A custom spam filter is developed to identify and eliminate these tweets based on this common characteristic.
2.3.2 Experiment set up

Terminology usage is modelled based on the input of all participants who correctly discuss the appropriate place/event, thereby showing a minimum threshold of interest and investment in proceedings. Therefore, participants are not just limited to those present at an event location (on-the-ground), but can also be those aware of and following the event.

Due to the globally high flow of tweets, the place/event combinations had to be carefully selected as to produce keyword searches that would produce a healthy amount of relevant material, without being drowned out by unrelated tweets. This can happen when keywords used have multiple meanings/uses or are too generic. Thus, places were selected based on the following criteria:

- Non-common terminology of location and event – Names that have few or no shared synonyms with existing words or places. Determined by cross-referencing with authoritative gazetteers (Committee for Geographical Names of Australasia, US Board of Geographic Names), alternative spatial databases (Geonames, Yellow Pages, OpenStreetMap) and general online sources (Wikipedia).
- High public interest events – Events that have significant hype and participation to generate high levels of Twitter content.
- High capacity location – Locations that can handle large numbers of active participants in order to increase the probability of obtaining on-the-ground tweets.

Based on these criteria, the following experiments are carried out (Table 2-1). Each place/event has varying capacities, duration, interest levels and activity type. Streaming is initiated at least three days prior to event occurrence in order to model the required stable profile of tweet rates.

2.3.3 Burst detection

The ability to discover a possible event in a monitored Twitter stream is achieved via statistically analysing the tweet rates of a known place name over time. With the assumption that tweet rates associated with these places will follow a normal distribution during non-event periods (discussed in Section 2.3), any deviation from a said normal profile can be seen as an anomaly and potential event. This method shares similarities to research that attempted burst detection in search engine queries for specific festivities-related keywords like “thanksgiving” and “Christmas gifts” (Vlachos et al. 2004).
Table 2-1: Data collection

<table>
<thead>
<tr>
<th>Official name</th>
<th>Type</th>
<th>Events</th>
<th>Collection period</th>
<th>Tweet Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flemington Racecourse</td>
<td>Outdoor racetrack</td>
<td>• AAMI Victoria Derby Day</td>
<td>6/11/2012 – 19/11/2012 (14 days)</td>
<td>64,000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Emirates Melbourne Cup Day</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Crown Oaks Day</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Emirates Stakes Day</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flemington Racecourse</td>
<td>Outdoor racetrack</td>
<td>• Big Day Out (Music fest)</td>
<td>25/1/2013 – 17/2/2013 (24 days)</td>
<td>900</td>
</tr>
<tr>
<td>KFC Yum! Centre</td>
<td>Multi-purpose arena</td>
<td>• Louisville basketball</td>
<td>30/11/2012 – 25/1/2013 (57 days)</td>
<td>204,600</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Dave Matthews Band</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• NCAA women’s volleyball</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mohegan Sun Arena</td>
<td>Multi-purpose arena</td>
<td>• One Direction</td>
<td>29/11/2012 – 25/1/2013 (58 days)</td>
<td>1,216,000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Smashing Pumpkins</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Dave Matthews</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Coldplay</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Quodrophenia</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Detection is done with the calibration of three parameters; (1) trailing moving average (TMA) window, (2) confidence threshold and (3) tweet threshold. Each parameter is adjusted in increments and in various combinations to each other to optimize for the best levels of precision and recall of each place/event. An event is said to be successfully detected if a burst is triggered within a 12-hour window before or after the official event start (time based on press-released data). Each detected anomaly is represented based on the hours of lag, post or prior to the known start time.

**Trailing moving average window**

The trailing moving average (TMA) window is used to systematically generate a future prediction of tweet rates for every hour based on an aggregated trend of past values. This is done to compensate for irregular tweet rate fluctuations not associated with a specific event. Among the major known patterns behind these fluctuations are the 24-hour tweet cycle model caused by the geographical distribution of users, time zone, and user habits. Intervals of a quarter of a day (6 hours) and half of a day (12 hours) were chosen to nullify this issue. In general the number of past values used will affect the sensitivity of burst detection. Hence, the smaller the window, the more responsive the system is to slight fluctuations, and the higher is the probability of triggering false positives, and vice versa.
Confidence threshold $\sigma$

With the TMA established, any tweet rate value falling outside this prediction is said to be an outcome of an extraordinary circumstance, which is speculated to be an event. The extent of how far this value has deviated is quantified using confidence thresholds. The confidence threshold is a general statistical method used to imply that the trend of tweet rates over time will still naturally fall within a normal distributed bell curve and any outlier is seen as a unique incident or event that is occurring. The system can be set to filter at one-, two- or three-times $\sigma$ (68%/95%/99%), which are the percent thresholds of the normal tweet rates versus strong outliers. The confidence levels are customized for each location to produce the best recall results.

Tweet threshold

The tweet threshold offers an additional quality failsafe for detection. This threshold refers to a pre-defined fixed minimum tweet count required to permit a burst detection to be triggered. This concept is implemented to compensate for the TMA window and confidence threshold that statistically perform poorly when dealing with very low tweet numbers (less than 10-20 tweets per hour). This minimum threshold is also useful to handle background noise that can skew detection and generate false positives when tweet rates are low. The value used is based on the average tweet rates recorded during non-event periods which reflects the statistical minimum stream rate for that place.

2.3.4 Event name discovery

Once a potential event has been detected, a sample of tweets from that triggered period is processed in order to extract an event name. If an event is actually occurring and the burst detection successfully identifies it, the tweets should contain a high occurrence of the event name too.

A semi-automated machine learning method is implemented to eliminate false positives (tweets of wrong context). This machine learning method functions on the basis that tweets relating to the same topic/subject matter would invariably contain very similar terms or phrases. For example, it can be expected that a discussion regarding a live concert would contain content related to the artists involved, location/venue, time of event, weather conditions, turn-out, atmosphere, etc. However, a user assigning pre-supposed words to be
used as a filter would generate a biased analysis. Hence, instead the user sifts through a random sample of tweets from the collected batch, reads the tweets as a whole, and selects/validates which of each actually refers to the appropriate topic. Validated tweets are then broken down into its individual tokens/words in order to determine occurrence levels.

For example, a random sample of two hundred tweets from the “Flemington Racetrack” stream (10% sample) was manually annotated and produced strong associated tokens of “racing”, “carnival”, “Melbourne”, and “Holiday”. These tokens are used as an automated filter to evaluate the entire dataset and eliminate inapplicable tweets. The tweets are cleaned of stop words, punctuation and common articles (e.g., hashtags, emoticons, aliases). The official place name is also removed. The tweets are then broken up to their individual tokens/words and ranked by frequency of occurrence. With the assumption that the official place name is removed along with all other common English language articles, the top occurring tokens left is assumed to potentially be the event name.

2.3.5 Substitution verification and monitoring
With the event name uncovered, its usage rate in tweets is compared to the usage rate of the official place name. Verification and monitoring of the discovered event name is carried out by initiating a new Twitter API stream based on each newly discovered term. The new event name terminology usage is compared with the existing official place name stream to determine which term is being used the most.

As discussed earlier, terminology substitution is said to occur when/if event name usage surpasses the official place name usage. As substitution is considered temporary, the event name lifespan ends once the relevant usage trend drops below the official place name usage. For now, the substitute term is known beforehand and is collected for post-processing, but this segment can be programmed to begin its own new API search in future work.

2.4 Experiment results
Section 2.4.1 shows a sample scenario of the Twitter streaming and monitoring process (before, during and after an event occurrence). Section 2.4.2 contains the tabulated burst detection results for two of the major test sites, Flemington Racecourse and KFC Yum! Centre.
Finally, Section 2.4.3 reveals the results of the semi-automated event name extraction process as triggered from burst detection.

### 2.4.1 Burst detection via data streaming

The result of burst detection is shown from the perspective of the data stream over time. Figure 2-1 depicts an example of detecting *Crown Oaks Day* event from the entire Melbourne Cup Carnival duration. This event is held at the Flemington Racetrack, a major horseracing venue in Melbourne, Australia. This particular carnival lasts for two weeks and consists of many race day events. Each race day itself is a sub-event with a unique name, usually with a sponsorship prefix. Actual on-the-ground attendance was approximately 100,000, and a total of 73,000 tweets were collected over a period of 14 days. Three sub-events are observed in this experiment, the *Emirates Melbourne Cup, Crown Oaks Day and Emirates Stakes Day*. Each of these sub-events occurred two days apart from each other. The official kickoff times of each event is marked by the blue dashed arrows in Figure 2-1. Thus, the example below looks specifically at detecting the middle event of *Crown Oaks Day*.

![Figure 2-1: Twitter stream of Crown Oaks Day verification. Search1 (blue): official place name usage, search3 (yellow): discovered place name substitute, TMA (black): normalized 6-hour tweet rate, BurstMark (red dot): triggered bursts potentially representing an event](image)

*Figure 2-1: Twitter stream of Crown Oaks Day verification. Search1 (blue): official place name usage, search3 (yellow): discovered place name substitute, TMA (black): normalized 6-hour tweet rate, BurstMark (red dot): triggered bursts potentially representing an event*
The primary tweet stream is shown as Search1 (blue) and represents tweets containing the official place name “Flemington Racecourse” and its common derivative “Flemington Racetrack”. The TMA line (black) is the normalized trend line of this stream based on a 6-hour moving window. The x-axis is the time progression in GMT hours and days (in brackets) from collection initiation. Day zero (0) represents the first day of streaming which was the 3rd of November 2013. The three prominent humps represent the spike in trend discussions generated from the three race-day events. The red dots are triggered bursts, which are potential (but unverified) events. From the perspective of Crown Oaks Day (middle hump), the closest detection to the actual event kick-off was on Day 5 at GMT 00:00 hours, which equates to a detection lag of -4 hours.

Once a burst is detected (red dots), the process of event name discovery takes place (results shown in Section 2.4.3) in order to extract the alternate event name label. In this case, the correct Oaks Day term is found. A new keyword tweet stream is then initiated based on this term (Search3, yellow) in order to verify its usage rate and compare it with the primary stream (Search1). As can be seen, the substitute term continues to stay relevant until the middle of Day 6. Once the new event name stream dips below the official place name usage, it is said that the event uptime is over and is no longer relevant enough to be defined as a substitute.

2.4.2 Overall burst detection results

Table 2-2 and Table 2-3 below are two examples of results from the collected tweet stream data for the Flemington Racetrack and KFC Yum! Centre. Depicted date and time info is based on the local time of event (Melbourne time zone offset of GMT +11). As shown in Table 2-2, a TMA window of six and twelve hours is tested for the Flemington Racetrack, each with 1σ and 2σ confidence levels. For Flemington Racetrack, confidence levels of 3σ were unusable and omitted due to the poor performance of detecting any of the relevant events, falling below a 50% success rate. As discussed in Section 4.3, the tweet threshold is calibrated based on the average down-time tweet rates, which in this case is ten tweets per hour. The event detection column represents incidents of recorded bursts and the temporal deviation of those bursts from the official kick-off time of each event. Negative values indicate detections hours prior to kick-off and positive values indicate detection during/after the event has started. As shown in the results, a single event can trigger multiple bursts throughout its lifespan. This represents an unstable twitter stream that is still fluctuating in relation to its general increased growth rate.
The other test site, the KFC Yum! Centre is a multi-purpose theatre in Kentucky, US. The place name, like many other arenas, is a product of its current sponsorship deal with the KFC fast food chain. It is not uncommon for venues such as this to continually and frequently undergo changes in sponsored naming rights (Rose-Redwood et al. 2008; Rose-Redwood 2011). The venue is the official home ground for the Louisville Cardinals basketball team and regularly hosts their games. Additionally, the venue can be rented for one-off events such as concerts or conferences. A period of one month streaming is carried out in order to obtain a dense period of both regular basketball games and unique events. Five of those events are shown in Table 2-3. Events 1 and 3 are the Cardinals basketball games, Event 2 was a concert by the Dave Matthews Band and Event 4 and Event 5 were the semis and finals of the NCAA Volleyball championships. The results were based on the same six and twelve hour TMA window but with the confidence threshold of 2σ and 3σ. The average tweet rate during the non-event period was 30 tweets per hour, so the tweet threshold is calibrated accordingly. The results from the 1σ setting produced poor precision and recall (less than 50%) and were omitted. Three out of the four parameter combinations failed to detect Event 5 (NCAA Volleyball championships) while all other events managed to be detected at least once.

Table 2-2: Flemington Racetrack detection results

<table>
<thead>
<tr>
<th>PARAMETERS</th>
<th>EVENT DETECTION</th>
<th>PERFORMANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMA window (hours)</td>
<td>Event 1</td>
<td>Event 2</td>
</tr>
<tr>
<td>Tweet threshold</td>
<td>Melbourne Cup</td>
<td>Oaks Day</td>
</tr>
<tr>
<td>Confidence threshold</td>
<td>6/11/2012 11:20</td>
<td>8/11/2012 15:40</td>
</tr>
<tr>
<td>6 10 σ 1</td>
<td>21 -9,-8,-6,-4,-1,0, +12</td>
<td>-17,-7,-5,-4,0, +12</td>
</tr>
<tr>
<td>6 10 σ 2</td>
<td>4 -9,+12</td>
<td>-7</td>
</tr>
<tr>
<td>12 10 σ 1</td>
<td>16 -9,-8,-6,-4,-1,0</td>
<td>-7,-5,-4,0</td>
</tr>
<tr>
<td>12 10 σ 2</td>
<td>13 -9,-8,-6,-4</td>
<td>-17,-7,-5,-4</td>
</tr>
</tbody>
</table>
Table 2-3: KFC Yum! Center detection results

<table>
<thead>
<tr>
<th>PARAMETERS</th>
<th>EVENT DETECTION</th>
<th>ANALYSIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Bursts</td>
<td>Event 1 Louisville Cardinals</td>
<td>Event 2 Dave Matthews Band</td>
</tr>
<tr>
<td>TMA window (hours)</td>
<td>Tweet threshold</td>
<td>Confidence threshold</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>σ 2</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>σ 3</td>
</tr>
<tr>
<td>12</td>
<td>30</td>
<td>σ 2</td>
</tr>
<tr>
<td>12</td>
<td>30</td>
<td>σ 3</td>
</tr>
</tbody>
</table>

2.4.3 Performance of event name discovery

Table 2-4 and Table 2-5 show the results of uncovering an event name via the semi-automated process for both sites. Tweets were taken from the most optimal burst point of each event (closest to kick-off time) and collected in hour batches. In order to train the context filter to remove non-related tweets, a sample of tweets is manually evaluated and indexed. Common tokens found in the approved tweets are used as guides to automatically include/exclude the entire dataset. An example of this is shown in Table 2-6. The filter then breaks down each approved tweet to its individual tokens, removes irrelevant noise (e.g., punctuation, symbols, and emoticons) and ranks tokens by occurrence. The assumption is that the high-ranking tokens should represent the event name, as seen in the underlined token results.
Table 2-4: Melbourne Race Carnival event name discovery

<table>
<thead>
<tr>
<th>Burst sample used</th>
<th>Melbourne Cup</th>
<th>Oaks Day</th>
<th>Stakes Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweet sample</td>
<td>GMT 19:00 (−9 hours prior event)</td>
<td>GMT: 00 (−4 hours prior event)</td>
<td>GMT: 6:00 (+2 hours during event)</td>
</tr>
<tr>
<td>Total tweets</td>
<td>3 hours</td>
<td>3 hours</td>
<td>3 hours</td>
</tr>
<tr>
<td>Tweet sample</td>
<td>590</td>
<td>870</td>
<td>42</td>
</tr>
<tr>
<td>Total tweets</td>
<td>per 10 tweets</td>
<td>per 10 tweets</td>
<td>per 1 tweet</td>
</tr>
<tr>
<td>Sampling interval</td>
<td>1134</td>
<td>1381</td>
<td>692</td>
</tr>
<tr>
<td>Total tokens</td>
<td>397</td>
<td>469</td>
<td>260</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Excluded stop words</th>
<th>Melbourne Cup</th>
<th>Oaks Day</th>
<th>Stakes Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 10 token results</td>
<td>89.8% cup</td>
<td>77.9% day</td>
<td>52.4% carnival</td>
</tr>
<tr>
<td></td>
<td>86.4% melbourne</td>
<td>76.7% oaks</td>
<td>31.0% spring</td>
</tr>
<tr>
<td></td>
<td>27.1% good</td>
<td>17.4% crown</td>
<td>23.8% day</td>
</tr>
<tr>
<td></td>
<td>25.4% day</td>
<td>12.8% oaksday</td>
<td>19.0% melbourne</td>
</tr>
<tr>
<td></td>
<td>15.3% going</td>
<td>9.3% carnival</td>
<td>16.7% cup</td>
</tr>
<tr>
<td></td>
<td>13.6% places</td>
<td>8.1% today</td>
<td>14.3% 4</td>
</tr>
<tr>
<td></td>
<td>13.6% firm</td>
<td>7.0% racing</td>
<td>14.3% \u201ctw_bet</td>
</tr>
<tr>
<td></td>
<td>11.9% first</td>
<td>7.0% good</td>
<td>14.3% race</td>
</tr>
<tr>
<td></td>
<td>11.9% corneliusracing</td>
<td>7.0% coverage</td>
<td>11.9% over</td>
</tr>
<tr>
<td></td>
<td>10.2% flemington's</td>
<td>5.8% theage</td>
<td>11.9% de</td>
</tr>
</tbody>
</table>

Table 2-5: KFC Yum! Center event name discovery

<table>
<thead>
<tr>
<th>Burst sample used</th>
<th>Louisville Cardinals (Basketball seasonal game)</th>
<th>Dave Matthews Band (Live concert)</th>
<th>NCAA Women’s Championship (Volleyball Tournament)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweet sample</td>
<td>GMT: 16:00 (+8 during event)</td>
<td>GMT: 23:00 (+9 during event)</td>
<td>GMT: 19:00 (+5 during event)</td>
</tr>
<tr>
<td>Total tweets</td>
<td>3 hours</td>
<td>3 hours</td>
<td>4 hours</td>
</tr>
<tr>
<td>Tweet sample</td>
<td>490</td>
<td>590</td>
<td>58</td>
</tr>
<tr>
<td>Total tokens</td>
<td>10</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Sampling interval</td>
<td>796</td>
<td>900</td>
<td>956</td>
</tr>
<tr>
<td>Total tokens</td>
<td>360</td>
<td>395</td>
<td>435</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Excluded stop words</th>
<th>Louisville Cardinals (Basketball seasonal game)</th>
<th>Dave Matthews Band (Live concert)</th>
<th>NCAA Women’s Championship (Volleyball Tournament)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 10 token results</td>
<td>34.7% louisville</td>
<td>42.4% dave</td>
<td>20.7% volleyball</td>
</tr>
<tr>
<td></td>
<td>24.5% illinois</td>
<td>42.4% matthews</td>
<td>20.7% championship</td>
</tr>
<tr>
<td></td>
<td>20.4% state</td>
<td>40.7% band</td>
<td>17.2% ncaa</td>
</tr>
<tr>
<td></td>
<td>16.3% vs</td>
<td>33.9% cliff</td>
<td>13.8% pm</td>
</tr>
<tr>
<td></td>
<td>16.3% game</td>
<td>33.9% jimmy</td>
<td>13.8% welcome</td>
</tr>
<tr>
<td></td>
<td>14.3% cards</td>
<td>28.8% w/</td>
<td>12.1% texas</td>
</tr>
<tr>
<td></td>
<td>14.3% love</td>
<td>23.7% others</td>
<td>12.1% carpet</td>
</tr>
<tr>
<td></td>
<td>14.3% go</td>
<td>22.0% dmb</td>
<td>12.1% arena</td>
</tr>
<tr>
<td></td>
<td>12.2% streak</td>
<td>18.6% louisville</td>
<td>12.1% out</td>
</tr>
<tr>
<td></td>
<td>12.2% center’s</td>
<td>15.3% ky</td>
<td>12.1% ncaavb</td>
</tr>
</tbody>
</table>
Table 2-6: Examples of context filters from manual annotation used for machine learning

<table>
<thead>
<tr>
<th>Tweet stream</th>
<th>Context tokens (for machine learning)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flemington Racetrack</td>
<td>Melbourne, Carnival, racing, festival, horse, winner</td>
</tr>
</tbody>
</table>

2.5 Discussion of detection results, areas of refinement and limitations

The overall event detection performance is discussed (Section 2.5.1) followed by a more detailed look at the effect of the parameters and settings on the results (Section 2.5.2). Section 2.5.3 looks at the event name discovery process and the NLP challenges that arise due to the medium and users involved. Finally, Section 2.5.4 provides a theoretical discussion of the accuracy of this model and the extent it represents actual place name usage in a society.

2.5.1 Overall event detection performance

Table 2-2 and Table 2-3 show the extent of detection abilities based on a variation of parameters and event type. For the Flemington Racetrack, all three event days were successfully detected with very similar detection times that remained consistent regardless of parameter combinations. Each event had a common detection point of the following; Event 1 (Melbourne Cup) at -9 hours, Event 2 (Crown Oaks Day) at -7 hours and Event 3 (Emirates Stakes Day) at -4 hours. These results reveal a continual detection time reduction of 2-3 hours per subsequent event. The consistency of detection times regardless of parameter combination shows an evident fixed point of a discussion beginning. There was also high detection redundancy, with an average of 3.8 burst incidents per event.

Meanwhile the KFC Yum! Arena produced trigger points that were temporally much further from the actual event, either being much earlier or later. Additionally, there was a failure to detect Event 5 for all but one parameter combination, resulting in a much lower recall. The average detection times were 16.7 hours prior event and 11.0 hours post event kick-off, regardless of parameter settings utilized. Event 2 (Dave Matthews Band concert), which was scheduled at 7pm local time, triggered a burst almost exactly one night before the event itself, while Event 1 (Cardinals Basketball league game) achieved a burst nine hours before its scheduled 1pm game.
This reveals that the Melbourne racing carnival had closer-to-event triggers compared to the KFC Yum! Centre and its respective events. It is speculated this is influenced by two major factors; the participation levels of non-present users and the nature of the event content itself. Melbourne Carnival is a globally followed horseracing event, which involves the activity of betting and monitoring race results. The higher stakes involved with this event can create an invested ongoing interest for a larger scope of people. Therefore, the frequent burst could result from discussions about the upcoming or finished races throughout the day. On the other hand, events of a predefined nature (concerts, theatre, etc.) are of a lesser relevance to non-attendees and possibly generate interest only a day before or after, since there is less immediacy to share the event moments. These events are also less likely to be broadcast live, hence only those in attendance can seed any authentic thoughts of a running event. This is evident in the KFC Yum! Arena, where the aired Cardinals basketball game had more responsive detection times than the concerts.

Another scenario to consider is when an event is composed of sub-events spread over a few days. In the case of the Melbourne races, the more prominently known Melbourne Cup day significantly overshadows the Crown Oaks Day on race-day two and Stakes Day on race-day three. The amount of back-chatter created from the “Flemington Racetrack” keyword stream for day one overshadowed the subsequent days. Fortunately, the right combination of parameters and the sufficient time interval between each event allowed all three to be successfully detected (see Section 6.2 regarding parameters). Discussion overshadowing is less of an issue for sub-events that have a more stable level of public exposure. For example, the Cardinals basketball matches that are scheduled over an entire league season showed consistent tweet rates for the observation period, specifically for Events 1 and 3 of the KFC Yum! Centre. Both events had the exact same burst time regardless of parameters utilized. Event 1 (game 1) had a -9 hour detection and Event 3 (game 2) had a -5 hour detection. This supposedly can be assigned to the relatively stable and unchanging fan base generating the majority of discussions.

2.5.2 Analysis of parameter modelling techniques
The tweaking of parameter settings allows for the possibility of moulding detection systems to compensate for event variations and characteristics. The following discusses the possible effects of these combinations and the reasoning behind them.
While each place-event combination is unique, there are common parameters that responded relatively well across the board for burst detection. In general, outside of events, tweet streams peak and dip on a daily basis based on daylight hours and also weekday/weekend cycles (Vlachos et al. 2004). This fluctuation is usually tied to the time zone of the place-event in question. Utilizing a six or 12-hour moving average window allows for the neutralization of these general fluctuations and seems to work well in practice. For both the Flemington Racetrack and KFC Yum Centre, increasing the TMA window size did not reduce the timeliness of detected burst and, in fact for Flemington Racetrack, it actually lessened the gap. Overall a 12-hour window produced the best precision and recall results for KFC.

Confidence thresholds represent detection sensitivity. The thresholds had to be customized based on the prominence and type of event in question, making it hard to standardize without prior experience. For Flemington racecourse, when fixing the confidence threshold at 2σ, the wider 12-hour TMA window resulted in bursts closer to the actual kick-off time. However, the most balanced results were from the six-hour TMA window and 2σ confidence threshold combination that yielded a perfect precision and recall. Overall, varying the parameter settings had more impact on precision and recall rather than the promptness of the detection itself. The temporal factor is very much conceded to the collective behaviour of the Twitter users and when they choose to trend a topic.

Tweet thresholds are harder to standardize because each place-event will not have similar interests levels and public participation. For example, Flemington had a tweet range between 60 and 4600, while KFC Yum! Centre had a range of 63 to 1001, even though both had similar average tweet rates during their stable periods. It is still essential to set a minimum threshold based on a profiled non-event streaming period to eliminate the skewing of detection. This skewing occurs when statistical methods are applied to extremely small data amounts. Future work could attempt to solve this issue by either employing system learning or by creating specific profiles based on venue details.

Overall, in terms of data quality, the more tweets, the higher the redundancy factor, which can equate to less noise and higher accuracy. However, this effect can be hampered if spam bots are not eliminated. Global events like Melbourne Cup had a spam levels of 19.1% (12,555 tweets), while the KFC Yum! Centre had a spam rate of 7.0% (10,989 tweets) based on a 0.7
followers/following threshold and 800 user following limit. Due to their nature, spam rates continually scale with tweet rates and can heavily distort datasets.

The influence of false positives is also considered, which are discussions that spike outside the accepted event time window. These occur when tweet streams for place/events are volatile and can be associated to general discussion tweets and re-tweeting regarding the event. For example, Event 2 at the KFC Yum! Centre (Dave Matthews Band concert) created post-chatter spikes up to two days after the event. Less prominent events like the Crown Oaks Day and Stakes Day at the Flemington Racetrack benefit from a lower confidence threshold due to the smaller and less exponential spikes of tweets. However, there were significant false positives for all but the 6-hour TMA/2σ parameter combination. Expanding the TMA window caused a decrease in all detected bursts, both the positives and false-positives.

2.5.3 Challenges in natural language and Twitter talk
Event name discovery proved relatively successful apart from minor flaws in phrase completeness (Table 2-4 and Table 2-5). Based purely on simple NLP machine learning and token ranking, the implemented system managed to uncover all of the event name terminology except for Oaks Day. Melbourne Cup and Oaks Day had a high degree of contrast between their phrases and the extracted next most occurring tokens, with a frequency above 85% and 75% respectively. The Dave Matthews Band concert and NCAA Volleyball tournament were also the top ranked phrases but with a lesser degree of ranking contrast, respectively generating a 6.8% and 3.4% jump above its subsequent token. Once the ranking contrast begins to dissipate, the task of distinguishing event terminology becomes harder. The event name discovery for the Cardinals basketball game was successful, but lacking in completeness. This could be attributed to the event itself being more descriptively varied in nature. For example, it can be associated with the team nickname (Cardinals), home ground state (Illinois), league association (NBA playoffs), match details (Cardinals vs. Hornets) and generic description (men’s basketball game).

Another major challenge is dealing with user-generated aliases for place names and event names through mash-ups, abbreviations, and alterations. This is common in natural language and more so in tweets, whereby character length is limited per tweet and there is a motivation to be syntactically prudent. For example, the collective desire for self-organization by Twitter users results in topics being grouped and narrowed down to a few hashtagged (#) terms. This
self-regulatory system practiced by the Twitter community eventually results in a settling for one or two terms per topic. Thus, Dave Matthews Band is aliased to its acronym #DMB, the Emirates Melbourne Cup is shortened to #Melbcup, and Crown Oaks Day is easily shorted to #Oaksday via spacing removal. Nicknames also further confuse event name discovery. The Cardinals team is sometimes referred to as the “cards” and is the sixth most highly occurring token (Table 2-5). Basic acronyms were used in the NCCA Volleyball Championship (“ncaavb”) but at much lower usage rates. Thus, there is a need to intelligently recognize and differentiate between a new term and a known term that has been modified by the public, although the complexity and boundless possibilities of phrasing variations make this task difficult. A possible method to accomplish this is by utilizing the current place-event detection system but continue to branch out from the event name to find aliases. In theory, a known terminology should always provide a lead to uncover a subsequent evolution of a newer term. However, the affect of normalizing language too much or too little using NLP methods can sometimes dilute the original user intent (Eisenstein 2013), hence care must be taken when interpreting phrases.

Due to the constraints of Twitter streaming parameters (keyword search) and the multiple usages/associations of English words, it is also difficult to mine the intended tweets and phrases, while avoiding large amounts of irrelevant discussions. In this experiment, the target sites were checked and tested based on uniqueness of name. This is a luxury that cannot be side stepped if a real-world system is to be developed. To implement at a larger scale the user intent has to be further clarified via reconciling of location or context of tweets to provide a stronger verification of tweet relevance.

Another challenge is determining the phrase ordering of event names. The token ranking step disintegrates all word ordering, leaving no structure guide. For example, in the Cardinals basketball game, the resulting event tokens by descending frequency were “Louisville”, “Illinois”, “state” and “vs” (Table 2-5). Without contextual knowledge, there would be uncertainty of the extent of the phrases and how they were commonly sequenced in tweets. There is also a need to distinguish between common English grammar terms (i.e., subject, verb, predicate) and place names. Thus the event name “The Melbourne Cup” contains a common article, a popular city name and an object, but could be misinterpreted in multiple ways.
2.5.4 Social modelling – accuracy and implications

As responsive as a detection system can be designed, there is also a need to consider how well the produced results holistically represent terminology usage within a society. The first major issue in contention is the social reach of Twitter in modelling a society. As discussed in Section 2.1.5, social media usage can be skewed in terms of contributors, as far as demographics, socio-economic background, and geographic regions are concerned. For example, Twitter is known for being highly adopted among the teen to young adult age range, mostly in city and suburbs (Mislove et al. 2011; Liao et al. 2012). On the positive side, the activities modelled for this research topic are primed for these demographics. Inversely, this system would be less applicable for events being held in rural areas with limited internet infrastructure or social media users.

Additionally, choice of terminology can be influenced by communication intent and targeted audience of the user. Is the user more interested in having personal discussions with friends, re-tweeting other content, or providing original thoughts and alerts to the masses? In any case, these factors influence the style of writing and hence the terminology used. For example, a re-tweet will always copy exact terminology without much personal perspective from the user (Liao et al. 2012), a formal announcement would likely use more official terminology and a casual chat among friends is likely to include social vernacular (lingo).

Terminology variation can also result from the shifting of discussion between an online/digital and real-world setting. Users could either opt for standardized phrasing for both platforms or hold two distinct styles. Research on the habits of students communicating via short message service (SMS) reveal they are capable of picking and choosing terminology from an entire repertoire of the language and actively do so in their daily communication (Tagliamonte and Denis 2008). This research revealed the stronger influence was not necessary depth of vocabulary but rather personal choice. There is also a third possibility of both platforms mutually influencing each other via crossovers of terminology, hence further blurring lines of distinction between medium. This can already be seen in the prevalence of SMS lingo utilized in real-world discussions like “lol” (Laugh out loud), “btw” (by the way) and “FYI” (For your information) (Crystal 2001; Randall 2002). Thus, it is not far-fetched to see this communication behaviour extending to Twitter.
Finally, there is an issue of handling global events running simultaneously (intentionally or incidentally) with a similar event name. This would not affect burst detection that is solely dependent on the initial official place name, but would hamper the following step of validating the event lifespan. Difficulty in resolving the location of tweets results in difficulty in ensuring the context of the correct place/event pairing. An alternate but related challenge would be handling main events in a single location whose sub-events run simultaneously but at different sections of the venue. This is common for music festivals like alternative rock fest Lollapalooza and Australia’s Big Day Out that has an overall event name, but consists of many bands/artists performing concurrently at various stages.

In contrast to past works in generic event detection using social media (Section 2.1.5 - 2.1.6), this system allowed for specific, previously unknown terminology to be uncovered. The terms were directly associated with the official place name, and an observable overlap and switch between the official and alternative is observed. This is seen via the observed discussions that initially used official place names exclusively, then evolving to the use of both the official place name and alternative name, and finally the alternative place name being used independently. Furthermore the terminology uncovered would represent a high-usage (hence most applicable) alternative place name, since data is based on actual discussions among Twitter users. This is opposed to other datasets like Geonames that depend on a handful of subjective volunteers uploading new information. There might also be additional limitations to detection when a location and event share the same name. For example, the Glastonbury Festival in England is named after the region it is held. However this issue is seen as less significant to solve since the similarly of location with event would mean that there is no missing alternative name to be found for place name databases.
In the 1830s, during the early developments of Melbourne, Sir Richard Bourke was appointed the task to name most of the streets in the still existing city grid. Included in that grid is the self-named Bourke Street, which remains unchanged to this day. However, when communicating locations/directions to people or natural-language enabled machines, we must be mindful to not be confused with the six other Bourke streets less than a 15 minute drive away nor to mistakenly assume them for the deceptively similar Burke Court, Birk Road and Berg Avenue scattered around the city and neighbouring suburbs.

Street name duplication or pronunciation similarity is an old and persisting problem for most urban settings (Coveyduck et al. 2004; Ciccone 2012; Flowers 2012). It is one of the many types of place names that undergo problems due to clashes between two or more similar sounding names. This includes duplication of city names and regions within the same country and sometimes even the same state (Daily Mail Reporter 2008; Editorial 2014). For street names, relevant authorities attempt to maintain order by keeping them as few and spread out as possible via the implementation of relevant policies and regulations. Adherence to these new rules requires the appropriate tools capable of uncovering and mitigating these potential street name pronunciation clashes.

A common method for discovering similar sounding names is by employing the Soundex phonetic algorithm. It is one of the popular algorithms for this task and has become a default addition to many database systems. This algorithm has been applied to datasets relating to the administering of street names. For example, the state government of Victoria (Australia) provides an online search engine with the integrated algorithm for municipalities to crosscheck new or existing names for potential clashes. However, the Soundex system has well documented obvious flaws in quality (Patman and Shaefer 2001). It also lacks the ability to provide a pronunciation “closeness” rating of one street with another, meaning streets names are either deemed fully matching sound-wise or are totally rejected.
Additionally, when considering that these names represent a street network spread out over a region, then a spatial element comes into play. This is because the probability of confusion is affected by how physically close conflicting streets are. Thus, there is value in interpreting them as both a function of pronunciation similarity and physical space.

It is hypothesized that the quantification of street confusion and its possible spatial attributes for an urban environment by Soundex can be significantly improved. Based on research of existing phonetic and string matching algorithms, along with the considerations of the unique qualities of street names, the aim is to develop the most appropriate method of gauging similarity levels of street names. The developed algorithm is then applied to a real street network in order to uncover the spatial aspect of the street name conflicts.

Through this process, the following issues will be confronted:

- The types of conflicting street names
  - What are the similarities of high/low conflict streets?
  - What trends of phonetic sequences are there?

- Distribution of conflicting street names
  - Is there evidence of specific patterns or regional trends?
  - To what extent does street density affect the conflict clustering?

At an explicit level, this study will aid in the process of administering new street names by providing knowledge of which areas are prone to naming conflicts and what names or phonetics sequences to avoid per region. At a larger scope, the results would also be a direct evaluation of the overall street naming problem as benchmarked to state regulations of the area of study (Victoria). This will answer the question of how well the state government policies have been implemented.

In the long run, if further developed, the outcome of this research topic could help improve services needing location information via verbal communication. For example, dispatchers can be made aware of which combinations of areas and street names require higher/lower levels of detail in order to confidently fix a person to a spot. Therefore, an emergency call with the verbal cue “its total chaos down here at Bourke Street” would require more accompanying detail compared to the more perceptually unique “I just got robbed at Bionic Ear Lane. Please send help” (actual Melbourne street names).
3.1 Literature Review on street name similarity and efforts to mitigate them

Section 3.1.1 to 3.1.3 provides a brief background of street name similarity confusion, its implications and initial steps in confronting it. Section 3.1.4 to Section 3.1.8 then introduces the common algorithms utilized to uncover equivalency for words and names along with the latest research in combining different types of algorithms. Finally, Section 3.1.9 discusses some unique qualities of street names to consider when improving on current algorithms.

3.1.1 The role of street names in an organized urban environment
The modern western address system can generally be viewed as a three-tiered granularity locating method; an administrative locality (city/town/suburb), a street, and a unit (building/house). Most countries follow this convention, with minor variations in terms of granularity order (e.g., Japan) and possible additions of more sub-region classes (e.g., China, Russia). Administrative regions are by default large, more prominently known and of lower quantities per unit area. On the other hand, units are usually identified by a sequential numerical system. The value of the units is distinct/absolute and reduces possibilities of misinterpretation. However, they are not unique and are repeated frequently in each neighbourhood. Between the more popularly known, low granularity administrative regions to the more distinctly accurate, but repetitive unit number lays the critical street information. Thus streets can be seen as the important binder that links a more familiar larger space with a distinct, but highly duplicated numbered point.

The accessibility of street networks in terms of scale and granularity can lead to street names being used as a location identifier. Volunteered geographical information (VGI) experiments have shown the prevalence of using street names as a primary geospatial expression (Tytyk and Baldwin 2012). Thus it is not uncommon to hear directions being given purely as a function of two intersecting streets, such as “I'll meet you at the corner of Victoria Street and Hoddle Street”. The critical role they play in the binding of an address means that the ability to distinguish between street names will highly impact the performance of successfully resolving location. However, resolving street identity in daily communication can be hampered by the existence of duplicate or similarly sounding names.
Many major cities like Boston (US), Toronto (Canada) and London (UK) began from a less formalized plan or structure, usually growing organically as various industries and land needs arose (Southworth and Owens 1993). As urban development continued laterally, frequent merging or absorptions occurred between larger and smaller regions, resulting in a corpus of duplicated names from joining territories. For example, Budapest (Hungary) was the joining of two cities, Buda and Pest, in 1873, with each original city having its own corpus of street names. Governing bodies, aware of this scenario, began to understand the value of an organized addressing system (including street naming) and proceeded to set out guidelines to do so (Corwin 1978).

More recent cities are built with a long-term perspective and better urban planning, thus mitigating the effects of accidental street name duplication. One method this is accomplished is by using a grid-based system, as commonly found in North American cities (e.g., Philadelphia, New York City, Washington, D.C.), whereby streets would be built at right angles from each other and the street names themselves would follow a numbering sequence (e.g., First avenue, Second avenue, etc.) or letter sequence (Rybczynski 1996). Hence we could direct someone to “the corner of sixth and twenty-third”. However, the grid system is mostly confined to city centres because extending usage to the entire region would either lead to non user-friendly street names (e.g., one thousand, five hundred and fifty two street) or necessitate secondary address information (suburb/postcode).

Regardless of the type of street-naming system, there is also the issue of street names not being duplicates but still being very similar in pronunciation. The level of confusion may vary, but it still poses a challenge in properly resolving the streets when communication is verbal and not written. This leads to the current day scenario whereby large parts of the population are living in confined regions that contain a myriad of duplicates or very similar street names.

3.1.2 Implications of street name confusion
The impact of street name miss-communication is more than trivial and lead to serious repercussions for those residing in these related localities. There have been real life incidents of emergency responders being delayed or misdirected due to a mismatch of duplicate street names that existed on opposing sides of town (Toronto Star, 2006; Flowers 2012). There has also been documented loss of life for the same reasons (Ciccone 2012). Other critical areas
that are affected are delivery services, utility management (water, gas and electricity), traffic safety (disorientated drivers), and legal document consistency (Corwin 1978).

The secondary safeguards in resolving location in urban environments are also susceptible to failure. For example, additional address information might not be known or missing and GPS technology can be unreliable or unavailable (Ciccone 2012). At a practical level, a passer-by witnessing an accident should be able to effectively call emergency services and quote a nearby street name sign to disambiguate his/her location without having to provide too much supplementary information.

3.1.3 Action plans and challenges to implementing solutions
Major cities in the province of Ontario and Quebec (Canada) underwent a major clean-up of street name duplicates mainly caused by the amalgamation of different territories into bigger regions (Coveyduck et al. 2004). Bell Canada, the centralized 9-1-1 emergency handler for these cities, cited the motivation for this clean-up:

*Bell’s Canada’s 9-1-1 system is built on criteria common to that throughout North America and each address within a municipality must be unique. Neither computer system nor emergency call takers can distinguish between two identical addresses within a municipal entity.* (Coveyduck, Lee & Holiday 2004, Page 12)

In Australia, the Victorian Government Guidelines for Geographic Names 2010 implements a street naming policy whereby no street name pair can be equivalent (in spelling and pronunciation) based on a function of distance and region type. The limits set are 5km for metropolitan, 10km for suburbs and 15km for rural areas (Department of Sustainability and Environment, 2013). An online search engine tool named VICNAMES was developed by the Victorian Department of Sustainability & Environment in order to facilitate the administrating of this tool. However, since these policies were only established post city development, there are still many existing scenarios that fail these criteria and remain unchanged.

Among the challenges of eliminating duplicates are the legalities and signage costs, reluctant citizens and business wishing to maintain status quo, and maintaining consistency of knowledge in produced maps (Coveyduck et al. 2004). These challenges are bureaucratic in
nature and can be resolved in time, though it admittedly will be a slow process of years. The same hurdles would apply for similar sounding street names if they were to be renamed.

Regardless of the level of street name conflict and the effectiveness of policies involved, what is evident is the importance of street identity in determining location and the real-world fallouts when they are unsuccessfully resolved. Thus, there is a motivation to be able to actively and retrospectively monitor the situation as seen from a phonetic and spatial perspective.

3.1.4 Soundex and Phonetic similarity

Every place name has an original, intended manner of pronunciation based on its designator or caretaker. The character sequences of a name tie to a specific intended sound. However, confusion in communicating a place name is common and can come down to mistakes caused by:

- Pronunciation of the speaker (text to sound)
- Audible Interpretation of the listener (sound to text)

The discrepancies or errors can be caused by one or both parties. Regardless of the source of error, there are certain patterns of spelling that have a higher tendency of generating similar sounds. Based on the pronunciation or phonetics of a word it is possible to identify possible conflicts that might occur for it.

Russell’s Soundex (Russell 1918) is one of the earliest phonetic algorithms which attempted to evaluate words, specifically names, based on their pronunciation. In human articulation, different sounds are created by varying the locations and methods with which to restrict airflow in our vocal tract. For example, pressing both lips together (Bilabial region) while pushing air out from the air cavities creates the [b] sound. The Russell phonetic system is based on the concept that consonants generated in similar/close-by regions of our vocal tract should also sound similar. For example, labiodentals which are produced by touching the bottom lip to the upper teeth contain the similar consonants of [f] and [v].

Soundex is a default addition for many popular database software ('Soundex' 2013) and is also the basis of the current VICNAMES service for street name management (Department of Sustainability & Environment, 2011). However, later research revealed a consensus of the
obvious limitations of Soundex (Christian 1998; Patman and Shaefer 2001; Holmes and McCabe 2002), which generate an overall low precision (33%) and recall (25%). In addition, there is no gauge of match “closeness” or ranking available since the algorithm produces a single discrete code per name, leading to binary results. Each name is either an exact match or mismatch to every other name.

3.1.5 Limitations of Soundex

In the original Soundex algorithm, each name is represented by a four-character code, consisting of a single letter, followed by three digits. The first letter of the code mimics the first letter of the name while the rest of the code is based on numerical grouping of the consonants. These groupings were created based on the assumed articulation similarity of the consonants. Thus, names with matching codes are supposedly similar in pronunciation (Table 3-1). All vowels except the first are ignored. For example, the name MacPhie and McVee would both generate the Soundex code M210, signifying they are equivalent. A more detailed process of the algorithm is described in literature (Patman and Shaefer 2001).

<table>
<thead>
<tr>
<th>Code</th>
<th>Letter Groups</th>
<th>Articulation Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B, F, P, V</td>
<td>Labial</td>
</tr>
<tr>
<td>2</td>
<td>C, G, J, K, Q, S, X, Z</td>
<td>Gutterals and sibilants</td>
</tr>
<tr>
<td>3</td>
<td>D, T</td>
<td>Dental</td>
</tr>
<tr>
<td>4</td>
<td>L</td>
<td>Long liquid</td>
</tr>
<tr>
<td>5</td>
<td>M, N</td>
<td>Nasal</td>
</tr>
<tr>
<td>6</td>
<td>R</td>
<td>Short liquid</td>
</tr>
<tr>
<td>SKIP</td>
<td>A, E, H, I, O, U, W, Y</td>
<td></td>
</tr>
</tbody>
</table>

However, since its development, there has been well documented evidence of Soundex’ limitations, which result in an overall poor precision and recall performance (Christian 1998; Patman and Shaefer 2001; Holmes and McCabe 2002). This can be attributed to a few major causes as summarized in Table 3-2. As a whole, the consonant groupings are too generalized, making sweeping judgments of letter similarity. For example, there are seven letters assigned to be under code two (Table 3-1) including [S], [Z] and [G] despite the fact that they can sound very distinct from each other. Additionally, there is a lack of context interpretation in which some letters sound differently if paired with other letters, like the variation of [G] when by itself or in the form of [NG]. Also, the short code length of four characters means that any additional phonetic grouping needed for longer names are ignored, resulting in their phonetic tail end being disregarded. Finally, the rigidity of the assigned first letter rule means that all
equivalency potential would be ignored if the first letters of a name pair did not match. Hence, when implemented in database systems, the user is expected to sift through a list of produced “good enough” matches and handpick a desired entry even though many would be of low relevancy. The efficiency of this process is heavily dependent on the size of database and the commonness of the searched phonetic code.

Table 3-2: Summary of major flaws of Soundex

<table>
<thead>
<tr>
<th>Issue</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>First letter binding is too rigid</td>
<td>Ivan (1150) ≠ Aivan (A150)</td>
</tr>
<tr>
<td>Lack of contextualization of consonants</td>
<td>Pfeier (P116) ≠ Fiefer (F160)</td>
</tr>
<tr>
<td></td>
<td>*[p] should be silent when in context of [pf]</td>
</tr>
<tr>
<td></td>
<td>Wiltshire (W432) ≠ Wilshire (W426)</td>
</tr>
<tr>
<td></td>
<td>*The cluster of [ls] and [lsth] should be similar</td>
</tr>
<tr>
<td>Noise intolerance</td>
<td>Alfred (A416) ≠ Afired (A146)</td>
</tr>
<tr>
<td></td>
<td>*Data entry errors are not tolerated</td>
</tr>
<tr>
<td>Different transcription systems</td>
<td>Hsiao (H200) ≠ Xiao (X000)</td>
</tr>
<tr>
<td></td>
<td>*Non-roman language scripts may have multiple translating methods</td>
</tr>
<tr>
<td>Names containing particles</td>
<td>Alhameed (A453) ≠ Hameed (H530)</td>
</tr>
<tr>
<td></td>
<td>*Both can refer to the same person</td>
</tr>
<tr>
<td>Does not process long words</td>
<td>Gardenview (G635) = Gardendale (G635)</td>
</tr>
<tr>
<td></td>
<td>*code grouping only up to first 4 relevant characters</td>
</tr>
</tbody>
</table>

Additionally, since this topic deals specifically with street names, there is also an added need to deal with street suffixes. Soundex has no adequate way to deal with an extended name/word apart from either treating it as a single entity (in which case the four character code would be a severe limitation) or generating two codes per street name. Both scenarios would generate poor results. For example, the street name Victoria Lane and Victoria Avenue would both have the same Soundex code of V236. Hence, there existed a need to improve upon the current Soundex solution.

3.1.6 Alternative phonetic algorithms

Various alternative algorithms were created in an attempt to improve on Soundex performance. The concept of grouping consonants based on location of articulation was still a
major foundation, but many improvements and alterations were made. The Phonex algorithm solves some of the context issues of letter transformation and increases performance by 44% in finding true matches with a minor reduced accuracy by 0.2% (Lait and Randell 1996). In this system, consonants are divided into 8 groups (2 more than Soundex). Character transformation rules were applied before the coding step to account for the context of letters. For example, [ph] is converted to [f] if located at the beginning of a name to account for the variation in pronunciation. The research also claims to provide a more objective result evaluation method by benchmarking the algorithm with data from the Family History Knowledge UK, a list of equivalent surnames sourced from UK genealogy data.

Lawrence Phillips’s Metaphone (1990) is another algorithm that attempts to improve on Soundex (Snae 2007). It is a rule-based system that can be loosely described as a 16 group consonant category. The consonant groupings are relatively similar to previous versions except that the sound [th] was represented by [0] and [sh] represented by [x]. The subsequent follow-ups of Double Metaphone and the commercially distributed Metaphone 3-added more character transformation rules and has the ability to produce secondary and tertiary codes per name to account for name variants (if relevant). For example, Smith produces a code of SMX and SM0. Metaphone 3 claims to have a 99% success rate for English words and names common in the United States (Guy et al. 2012). As of yet, there has been no research on its performance with other culture/languages.

Attempts were also made to tackle specific cultures and languages that are ignored by the inherently English-centric algorithm of Soundex. NYSIIS (New York State Identification and Intelligence System) is an American adapted system used for US databases and records that could account for Hispanic names. It was abandoned in 1998 due to the much wider range of migrant ethnicities entering the US that did not fit well with the algorithm. Daitch–Mokotoff Soundex (Lait and Randell 1996) is a derivative of the original algorithm intended to compensate for the unique qualities of Jewish and Eastern European names. Caverphone (Hood 2002) addresses the name pronunciation and spelling of the Maori and was used to consolidate census data for the New Zealand elections. Each method to a larger extent succeeded in confronting their specific niche of culture/language. A summary of these algorithms is provided in Table 3-3.
### Table 3-3: Summary of prominent alternative phonetic algorithms

<table>
<thead>
<tr>
<th>Phonetic algorithm</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYSIIS (1970)</td>
<td>• American adapted system used for US databases and records that could account for Hispanic names.</td>
</tr>
<tr>
<td></td>
<td>• Abandoned in 1998.</td>
</tr>
<tr>
<td>Daitch–Mokotoff Soundex (1985)</td>
<td>• Soundex adaption for Slavic and German spelling of Jewish names.</td>
</tr>
<tr>
<td>Phonex (1996)</td>
<td>• Aimed at improving discovery of name equivalency based on genealogy records (not necessarily similar sounding names).</td>
</tr>
<tr>
<td>Caverphone (2002)</td>
<td>• Aimed at tackling Maori spelling of names in New Zealand.</td>
</tr>
<tr>
<td>Metaphone (1990)</td>
<td>• Developed by Lawrence Philips for matching words that sound alike.</td>
</tr>
<tr>
<td>Metaphone 3 (2009)</td>
<td>• Double Metaphone and Metaphone 3 are commercially available.</td>
</tr>
</tbody>
</table>

### 3.1.7 String similarity

Aside from phonetics, word and name similarity can also be evaluated based on string matching techniques. String matching evaluates one word with another, purely by their character content or sequencing. This research area has origins in the field of information retrieval that deals with efficiently searching for files and records from large database sets. Among the more popular methods are n-grams, Agrep (Wu and Manber 1992), Levenshtein/Edit Distance (Levenshtein 1966), and Longest Common Substring (Friedman and Sideli 1992).

The n-gram technique splits words into groups of neighbouring characters, usually done sequentially and with one character shift at a time. The number of characters per group denotes the n in n-grams. The character groups of two words are then compared for matches (Figure 3-1). The more number of common n-grams found in both words, the higher the similarity score.

| Jackson | [Ja],[ac],[cs],[so],[on] |

**Figure 3-1: Example of bigrams (n =2) which creates 5 groups of letters**
Agrep is an algorithm created to match a substring to a string database that differs by $k$ amount of insertions, deletions and replacements, whereby $k$ is a predefined constant (Zobel and Dart 1996). Results are not ranked and the system itself was not designed for phonetic matching but rather a quick method of searching many files.

Levenshtein/Edit Distance evaluates similarity of two words by counting the number of single character insertions, deletions and replacements needed to transform one string to the other with the assumption that words that are spelt relatively similarly should also sound similar. The Damerau–Levenshtein distance is similar except that it considers the transpositions of characters (Pfeifer et al. 1995).

The Longest Common Substring (LCS) attempts to evaluate similarity by finding the longest common sequence of letters in a word pair. The substring is then removed from both words, its length noted and the process looped until no common letters remain. This method was developed for the healthcare industry in order to find matching patient records.

### 3.1.8 Fusion methods

Fusion systems refer to methods that combine phonetic algorithms with alternative techniques, like string matching. Various papers (Pfeifer et al. 1995; Zobel and Dart 1996; Holmes and McCabe 2002; Snae 2007) testify to a significant increased performance in robustness and accuracy. It has the benefit of better results width that is usually missed by solely one method or the other. The healthcare sector, human resource management and criminal investigation are common intended benefactors of the research.

There was research in criminal investigations in order to identify criminals through their deceptive identities as recorded in police databases (Wang et al. 2004). Using a combined algorithm of Soundex and Edit Distance, they aimed to identify 24 criminals from a database of 1.3 million records, with each criminal having one authentic record and several fake ones. In addition to their name, the details of date of birth, residence and ID were evaluated. In accordance to police advice, each parameter was given equal importance and equal weight when calculating an overall score for a record. By varying the threshold of their algorithm, they were able to balance the accuracy, false positives and false negatives for optimal results.
In a purely performance driven endeavour there was an attempt to combine Soundex and Soundex adaptations with n-grams in order to increase the precision and recall of Soundex as a standalone (Holmes and McCabe 2002). A corpus of 90 names was chosen, each with judged collection of variant names (1187 in total). The algorithm acheived good results with this fusion method, at 96% precision.

The concept that a combination of techniques, common in the field of information retrieval, improves performance was also studied (Zobel and Dart 1996). The authors proposed Editex, which combines Edit Distance with phonetic based strategies from Soundex and Phonix. However, they also discovered that even high performance algorithms, when applied on the same dataset, can produce inconsistent results when compared with each other, almost without overlap. That meant each algorithm variant would produce different results. Therefore, it is important to evaluate the data, purpose and context on which an algorithm is to be used.

3.1.9 Differences of street names vs. generic names

Street names have certain characteristics that differentiate them from regular names, as analysed in most of the previously discussed research. For one, the etymology or variations of street names can be much more diverse. Recchia and Louwerse (2013) evaluated the performance of string similarity measures on place names based on variations in spelling, transliterations and unofficial labels. Toponyms along with their alternative labels were obtained from the GEOnet Names Server of the National Geospatial-Intelligence Agency (NGA) and run through an existing phonetic and string-based algorithm. They discovered that algorithm performance is extremely region/language dependent and the algorithms have to be benchmarked and customized to needs. Additionally, streets can posses a street suffix (i.e. Lane, Drive, and Avenue) for which the list of options is usually finite, but also possibly ambiguous in definition. Varol and Talburt (2011) looked at street name errors based on data entry mistakes using the Beider-Morse algorithm. This algorithm that had flexible rules that changed based on the detected language, hence making the processing style more adaptable. However, street suffix was not considered in this setup.

The methods for comparing algorithm quality are also different, considering the diverging goals of the previous works and this one. Most of the aforementioned research was directed towards solving name equivalency, especially in large citizen datasets. This was motivated by
the need to reconcile multiple databases of civilian information in various industries. On the other hand, this research topic of evaluating street name conflicts is purely a matter of confusion levels of one name with another, regardless of the genealogy or variation theories of names. Hence Derek Street, Tariq Street and Tei Lik Street should mutually trigger a similarity conflict with each other (with varying degrees of closeness) despite the fact they are not equivalent.

3.2 Designing a street name similarity algorithm

The aim of this topic is to find any possible relationship between pronunciation conflict and space, because the probability of confusing one street with another is tied to their regional proximity and level of clustering. Thus, to properly understand the conflict levels of a network of street names, the following is qualitatively evaluated:

- Components of street name similarity
- Phonetic and spatial trends of conflicting streets

Section 3.2.1 discusses the overall concept of quantifying street name similarity, which consists of an introduction to the separate components of the algorithm developed in this work (Section 3.2.1.1 to 3.2.1.5). Lastly, Section 3.2.2 discusses the spatial analyses methods implemented using the new algorithm.

3.2.1 Components of street name similarity

The user perception of a street name is seen as a function of pronunciation, spelling similarity, and street type/suffix. The more similar these three factors are for a tuple of streets, the higher the chance of user confusion. In the context of finding street name conflicts, the concept of ranking also plays a critical role because similarity levels exist on a continuous scale. The lack of ranking leaves authorities/policy makers in a difficult position of still having to sift through and interpreting approved results, which can be daunting if the dataset in question is large. Thus, having a scale of similarity provides a quantitative measure in which to base decisions.

Our proposed method attempts to improve on the popularly used Soundex algorithm by increasing precision/recall performance and providing ranking of results. This is achieved by
combining multiple evaluation methods into a single algorithm. The pronunciation component is handled by two phonetic-based algorithms, the existing Metaphone and our newly developed EndCode. Spelling/string similarity is carried out with n-gram analysis. Finally, the influence of street suffixes on street name confusion is conceptualized as a function of their frequency.

3.2.1.1 Algorithm component (1): Metaphone

Metaphone, a rule based phonetic algorithm, is utilized as the scoring backbone. It is chosen based on the evaluation of various open-sourced options due to its overall flexibility and performance levels. The Metaphone system showed the ability to properly handle mainstream contextualization better than Soundex (Lait and Randell 1996; Snae 2007) and relatively on par with the other available alternatives. Additionally, it is not rigid in rule-based character transformations. This attribute is important because the chosen phonetic algorithm is not the sole method for evaluating results, nor the only phonetics-based filter. Thus, it has to concede the more refined analysis to the remaining components instead of strictly eliminating all possibilities on the offset.

In Metaphone, pronunciation is divided into 16 different consonant sounds, as shown in Figure 3-2. Each letter of a word/name will be transformed to its equivalent consonant sound based on a set of predefined rules. These rules take into account the context of a letter and its neighbouring letters in order to produce a more accurate interpretation of its sound. For example, if the letter \([p]\) is followed by \([h]\), then both letters will instead be represented by the group \([f]\). This improves on Soundex by taking into account letter contextualization.

Some extra rules are added to refine the algorithm. The letter \([w]\) is dropped whenever not followed by a vowel or \([y]\) and is not the first or last character. The letter \([w]\) is seen as almost always being less influential to its neighbouring characters while not providing strong distinctions in sounds. Additionally, \([s]\) is disregarded at the end of a name if preceded by a consonant. This is to confront direct variations or very similar constructs of the same name (e.g., John/Johns).

![Figure 3-2: The 16 consonant sound categories of Metaphone](image-url)

<table>
<thead>
<tr>
<th>[B]</th>
<th>[X]</th>
<th>[S]</th>
<th>[K]</th>
<th>[J]</th>
<th>[T]</th>
<th>[F]</th>
<th>[H]</th>
<th>[L]</th>
<th>[M]</th>
<th>[N]</th>
<th>[P]</th>
<th>[R]</th>
<th>[O]</th>
<th>[W]</th>
<th>[Y]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero ([0]) represents the ([th]) sound</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.2.1.2 Algorithm component (2): EndCode

An additional phonetic component is developed to specifically address vowel sounds in names, called EndCode. Up to now, most phonetic solutions solely depended on grouping consonants to represent pronunciation, and vowels are only acknowledged when needed to contextualize their neighbouring consonants into their respective code groupings. While it is accepted that consonants are the dominating group that help anchor pronunciation, once they are modelled, there is a clear secondary influence of vowels (Ladefoged and Disner 2012). For example, while the average human can hear the difference in the name Packard and Piccard, most phonetic solutions would find them indistinguishable, including Metaphone (code PKRT) and Soundex (code P263). Even string matching methods like n-gram evaluation would observe a single character difference for both names (2nd letter), resulting in an identical (imperfect) score.

The concept of parameterizing vowels has mostly been shunned by traditional single-code solutions because it would increase the matching rigidness due to the addition of another grouping type. The outcome would be an algorithm of potentially better precision but much lower recall. Additionally, including vowel representation into the same code system would wrongly elevate vowels to the same influence level as the original 16 consonant groups. Algorithmically, they would also be difficulties in implementation because the vowels are intertwined within consonants with variable sequences and lengths, making them hard to isolate and group.

The better solution is to extract the vowels and evaluate them as a component of their own, thus not degrading/diluting the integrity of the original consonant-based code. This allows for a dual assessment of the phonetic qualities of a name, opening up the avenue for a partial phonetic match. This method shares some resemblance to Double Metaphone and Metaphone 3, which create secondary/alternate codes for a single name in order to increase chances of finding a match. However, these alternative codes (from Double Metaphone and Metaphone 3) are based on varying cultural character transformations and not an attempt on vowel representation.

At this stage the processing scope is limited to trailing vowels for two major reasons. Firstly, trailing vowels are much easier to parameterize compared to non-trailing vowels because the latter can be of more varied lengths, sequences and amounts when mixed with consonants in a name. Secondly, it is argued that the trailing vowels are more influential in the overall
perceptive shape of sound. Table 3-4 reveals examples of the stronger effect of trailing vowel sounds versus vowel sounds in alternative locations, towards providing a good representation of the uniqueness of a name.

**Table 3-4: Examples of variations in vowel sounds**

<table>
<thead>
<tr>
<th>Metaphone Code</th>
<th>Start/middle vowel variations</th>
<th>Trailing vowel variations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALNZ</td>
<td>Alonzo, Alonzo, Alenzo</td>
<td>Alonzo, Alonzei, Alonzia</td>
</tr>
<tr>
<td>AKSTN</td>
<td>Agustin, Agostin, Igasten</td>
<td>Agustin, Agustina, Agustinia</td>
</tr>
<tr>
<td>TRS</td>
<td>Durso, Dorso, Dirso</td>
<td>Durso, Dursi, Dursai</td>
</tr>
</tbody>
</table>

The proposed EndCode provides a supplemental numerical code to the phonetics. The code hasgrundings in the International Phonetic Alphabet (IPA), which is a standardized method of representing sounds in oral language (International Phonetic Association, 1999). The IPA categorizes vowels sounds based on the openness of the jaw and position of tongue of the speaker, as represented by each phonetic notation (Figure 3-3). The variations of both these elements are what generate all possible vowel sounds, regardless of the language/culture of a speaker; hence their representation is universal. In Figure 3-3, the vertical component represents the degree of jaw openness, while the horizontal component represents the position of the tongue (in relation to the mouth cavity). A more detailed explanation, along with audio clips that represent each phonetic notation can be found online (Wikipedia, 2014b).

Endcode works by categorizing the more similar IPA vowel coding into five groups. The associated letter combinations and code for each group are listed in Table 3-5. By taking the last two trailing letters of any name it is possible to model the final syllabic sound of a name. Thus, it is assumed any word/name sharing the same Endcode grouping should have similarity in trailing vowel sound.

The IPA categorizes vowels sounds based on the openness of the jaw and position of tongue of the speaker, as represented by each phonetic notation (Figure 3-3). The variations of both these elements are what generate all possible vowel sounds, regardless of the language/culture of a speaker; hence their representation is universal. In Figure 3-3, the vertical component represents the degree of jaw openness, while the horizontal component represents the position of the tongue (in relation to the mouth cavity). A more detailed explanation, along with audio clips that represent each phonetic notation can be found online (Wikipedia, 2014b).

Endcode works by categorizing the more similar IPA vowel coding into five groups. The associated letter combinations and code for each group are listed in Table 3-5. By taking the last two trailing letters of any name it is possible to model the final syllabic sound of a name. Thus, it is assumed any word/name sharing the same Endcode grouping should have similarity in trailing vowel sound.

![IPA diagram for vowel sounds](source: www.phonetics.ucla.edu)
Table 3-5: EndCode grouping

<table>
<thead>
<tr>
<th>Sound</th>
<th>Equivalent IPA code</th>
<th>Designated Code</th>
<th>Letter combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>“AH”</td>
<td>α, õ, ð, æ, e, ə,</td>
<td>1</td>
<td>[aa] [oa] [ua] [ia] [ah]</td>
</tr>
<tr>
<td></td>
<td>Λ, œ</td>
<td></td>
<td>Consonant + [a]</td>
</tr>
<tr>
<td>“AY”</td>
<td>æ, e, i</td>
<td>2</td>
<td>[ae] [ei] [aa] [ea] [ia] [eh]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[ay] [ey] [iy] [oy] [uy]</td>
</tr>
<tr>
<td>“EE”</td>
<td>e, i, y, i, i, ĩ</td>
<td>3</td>
<td>[ee] [ea] [ia] [ie] [ii] [oi]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[ay] [ey] [iv] [oy] [uy]</td>
</tr>
<tr>
<td></td>
<td>Consonant + [y]</td>
<td></td>
<td>Consonant + [i]</td>
</tr>
<tr>
<td>“AI”</td>
<td>ɪ, e, i</td>
<td>4</td>
<td>[ai] [aa] [ea] [ii] [oi] [ui]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[ay] [ey] [iv] [oy] [uy]</td>
</tr>
<tr>
<td></td>
<td>Consonant + [i]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>“OW”</td>
<td>θ, o, û, u, ŋ, u, D</td>
<td>5</td>
<td>[Ao] [au] [eo] [io] [ou] [oe] [uo] [ow]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Consonant = [o]</td>
</tr>
<tr>
<td>“OO”</td>
<td>o, θ, ũ, u, w, ɵ, e</td>
<td>6</td>
<td>[Oo] [iu] [ue] [uu] [eu]</td>
</tr>
<tr>
<td>Consonant or undetermined</td>
<td>- n/a -</td>
<td>0</td>
<td>consonant</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Consonant + [e]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(everything else)</td>
</tr>
</tbody>
</table>

Vowel type sounds can be generated from two consecutive vowels or a hybrid of consonants and vowels. The remaining consonants are considered to generate stop sounds caused by the restriction or closing of the vocal tracts and are assigned their own group together under code 0. A name can be assigned more than one code. Thus the name “Dorsey” with trailing letters [ey] would be assigned the code 2, 3 and 4. If a name pair has one or more EndCode groups that are the same, they potentially have similar trailing sounds and are considered a match. There are intentional overlaps in letter combinations, like [aa] which is associated to code 1, 2 and 4. This makes the solution more inclusive in nature by only filtering out clear mismatches. Again, by having multiple assessment techniques, each component can be afforded more flexibility.
3.2.1.3 Algorithm component (3): String matching

String matching is used to find any other possible character sequence correlation in a name pair. This method helps to compensate for variations of yet unknown or unclassified phonetic similarities (Donato 2008). Where the issues of noise intolerance, particle in names, and transcription variations stifle phonetics (summarized in Table 3-2), n-gram methods are still able to extract a degree of similarity purely based on sequences of spelling. Thus, if particles are added, like adding [AL] to change Hameed to Alhameed, n-grams would still recognise the remaining five bigrams that associate both names together. The assumption is that words sharing a high ratio of identical grams cannot stray too far in pronunciation too. The n-gram method not only looks at character similarity but the order they are assembled together. Therefore, to gain a high score a name pair has to not only share many common characters, but must be grouped similarly as well.

Using a larger n value increases the quality of matching grams (larger sequence matching) but also lowers the potential of finding gram associations. Based on previous research in information retrieval, bi-grams (two characters per gram) were chosen as the optimal choice for performance (Pfeifer et al. 1995). Additionally, the technique of white-space padding is used to provide additional weight to the first and last letters of a word. This is done by adding a blank character space before and after a word. Thus the name Jackson would have the additional bi-grams of [ _ j] and [n _ ] at the start and end of the name for matching. There are various equations used to produce a similarity score as discussed by Kondrak (2005) and Islam et al. (2012), and this research opts for a common method as utilized by Pfeifer (1995). Thus the coefficient of similarity (δ) is the product of association for name A with α n-grams and name B with β n-grams is as shown below:

\[
\delta = \frac{\alpha \cap \beta}{\alpha \cup \beta}
\]  

For example the street name Patinson with 9 bi-grams and Patterson with 10 bi-grams have a total of 6 common grams, thus generating a coefficient value of 0.46.
3.2.1.4 Algorithm component (4): Street suffix

Street suffixes/categories (i.e road, drive, avenue) are a common addition to street names and function as an additional descriptor of street function. It is common to find that street suffixes are often dropped, forgotten or mixed up in daily communication due to a lower threshold of importance that users assign to it. This confusion is further exacerbated when the various suffixes involved are less distinct in meaning or harder to differentiate like the collection of “Avenue”, “Lane” and “Grove”. It is also possible to find exceptions to the rule in which assigned suffixes do not fit the actual road function, like Park Lane being assigned to a highway. In summary, these suffixes can be a substantial source of confusion.

The number of street suffixes assigned to a region is usually much smaller than the amount of street names themselves, with a ratio of roughly 1 to 730 for Greater Melbourne. Furthermore, pre-test using Soundex and Metaphone reveal that all 88 suffixes in Greater Melbourne have zero phonetic correlation with each other. Hence, it can be seen that any confusion is less related to how similar they sound, but rather the potential of mixing-up the suffixes as a whole.

A simple solution is proposed, whereby the potential of confusing two different suffixes is related to how frequently it occurs within a network. This assumes that it is easier to mistake the most commonly used suffix of street with Road (3rd most frequent) as opposed to street with Drive (27th most frequent). Thus, from Equation 2, the street suffix score for two streets of suffix frequency $a$ and $b$ would be their combined influence as benchmarked to the total amount of street.

\[
\text{Street suffix score} = \frac{a + b}{c}
\]  

For example, based on the Greater Melbourne street network, the street suffix of street (frequency of 11,111) and road (frequency of 2943) would obtain a score of 0.24, which represents the probability of mistaking one with the other.
3.2.1.5 Final scoring

All four components are combined to produce a single algorithm. As this is an early stage algorithm implementation, each component is allocated equal weight on the total score. In the future, a more appropriate weighting solution may be necessary, depending on the results of this and similar research. The final algorithm produces a means to evaluate the similarity of two distinct streets by producing a score of match similarity between them, as shown in Equation 3.

\[
\text{similarity score for two streets} = [\text{Metaphone match}] + [\text{Endcode match}] + [\text{n-gram score}] + [\text{street suffix score}] \tag{3}
\]

A working example of the algorithm is shown in Table 3-6. Two street name pairs are shown in order to reveal some significant evaluation factors in the new algorithm. The first pair is *Boundary Road* and *Bindaree Court*. Both have matching Metaphone codes of BNTR and also trailing sounds of “EE” (EndCode 3) showing promising signs of similarity. However, they are not a perfect match, hence the addition of string evaluation (n-gram) and street suffix uncover these differences. The end result is an overall score of 0.64 (over 1). In the case of street C (*Jefers Street*) and D (*Cheffers Street*), there is a mismatch in the Metaphone, but a match for EndCode. However, the n-gram has a much stronger association (score of 0.46), hence lifting the overall final similarity score to 0.61. From these two basic examples, it can be seen how different aspects of similarity are taken into account.

**Table 3-6: Example of similarity scoring for two streets using the new algorithm**

<table>
<thead>
<tr>
<th>Phonetic Code</th>
<th>Metaphone</th>
<th>EndCode</th>
<th>n-grams</th>
<th>Street suffix</th>
<th>Full score (over 1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Street A: Boundary Road</td>
<td>BNTR 3</td>
<td>1.0 (match)</td>
<td>1.0 (match)</td>
<td>0.29 (partial match)</td>
<td>0.26 (partial match)</td>
</tr>
<tr>
<td>Street B: Bindaree Court</td>
<td>BNTR 3</td>
<td>1.0 (match)</td>
<td>1.0 (match)</td>
<td>0.26 (partial match)</td>
<td>0.61</td>
</tr>
<tr>
<td>Street C: Jefers Lane</td>
<td>JFRS 0</td>
<td>0.0 (no match)</td>
<td>1.0 (match)</td>
<td>0.46 (partial match)</td>
<td>1.00 (full match)</td>
</tr>
<tr>
<td>Street D: Cheffers Lane</td>
<td>XFRS 3</td>
<td>1.0 (match)</td>
<td>0.46 (partial match)</td>
<td>1.00 (full match)</td>
<td>0.61</td>
</tr>
</tbody>
</table>
3.2.2 Phonetic and spatial trends of conflicting streets

With the ability to quantify the similarity between two street names via the new algorithm, the subsequent step is to uncover spatial attributes in a street network. This is done by evaluating street pairs in groups based on their spatial proximity. Each street obtains a conflict score representing its level of perceptive similarity with all the streets surrounding it. The conflict score for Street A would be the sum of scores from its pairing with Street 1 to Street n, whereby n is the total number of streets in a defined area around Street A.

The extent of the street similarity conflicts are analysed at a global and local level. The global level represents the conflict of each street with the entire extent of the map area/database. This would reveal how distinct or generic a street name pronunciation is as benchmarked to the entire region. The local level represents conflict at a neighbourhood/ regional level by revealing clusters of streets that could potentially have pronunciation issues amongst each other. Each cluster represents a group of streets in close vicinity to each other and is grouped via a moving window. If the entire map were divided using an even grid system (Figure 3-4), each grid square would represent a region/cell that would inherit a discrete conflict value. The conflict value is obtained by overlaying a larger window directly on top of the cell and evaluating all streets contained or intersecting the window area. The sum of conflict scores for all streets processed in the window would be the conflict score of the cell. The outcome is a map with cell values for an entire map area that depicts a continuous view of the high and low areas of street conflict in the form of a heat map. The size of the cells and windows itself can be customized based on needs.

![Figure 3-4: Example of moving window concept for street similarity assessment in a region. The conflict value of the central cell is the combined conflict score of all streets intersecting the moving window.](image)

54
Clustering is also analysed from the perspective of administrative regions. Streets are collectively evaluated based on the municipality they belong to, with each municipality obtaining a single score representing its overall street name conflict levels. There is added significance to this analysis from the fact that the management of street naming is often in the authority of each municipality. Thus, the outcome represents the degree of compliance each municipality had achieved in adhering to the Victorian state policies.

3.3 Algorithm evaluation setup

Section 3.3.1 discusses the steps to evaluate the performance of the new algorithm and Soundex as benchmarked to a human-annotated set. The popular Levenshtein Distance algorithm is also included as a representative of a standard character-based scoring algorithm, used frequently as a benchmark for information retrieval research (Pfeifer et al. 1995; Zobel and Dart 1996; Snae 2007). The algorithms are analyzed from two perspectives, (1) the quality of results based on precision and recall; and (2) the quality of ranking. The tests are conducted using the Greater Melbourne street name dataset. Section 3.3.1 discusses the steps and parameters for this evaluation. Next, Section 3.3.2 applies the new algorithm on the same street name network in order to uncover spatial patterns and attributes in Greater Melbourne. The various variables involved in Greater Melbourne street names are discussed (Section 3.3.2.1 – Section 3.3.2.2), while the software and processing steps are discussed in Section 3.3.2.3.

3.3.1 Experiment design: Human-annotated test

In this experiment, a street name is randomly chosen from the Greater Melbourne street name network and run through the three algorithms (Soundex, Levenshtein Distance and the new algorithm) to find any possible matches with the entire Greater Melbourne street name corpus. The top 20 results produced by both the new algorithm and Levenshtein Distance along with all results from Soundex (Soundex cannot rank) are collated in a list for evaluation. Amazon Mechanical Turk (abbreviated to Mturk) is used as the source of human-annotated feedback for this list. Mturk is an online crowd-sourcing human intelligence marketplace, used by businesses or organizations to accomplish tasks that are difficult to automate by machine and require human decision-making skills (e.g., survey polls, image identification, transcribing audio recordings, etc).
A survey is created based on the pairing of the original street name with all its potential similar possibilities from the collated list (the list is randomized and cleaned of duplicates). Figure 3-5 shows a sample survey set for the street name “Diana Street” along with some possible matches. Participants are asked to imagine a hypothetical scenario of a verbal discussion between two native English speakers, and directed to rate the probability of them confusing one street name with another. Scoring options for each street pair are based on a scale of 0 (no probability) to 3 (high probability). This survey is replicated for a total of four sets of street names, each with 50 unique participants (200 surveys in total). Only participants with masters qualification are used. Masters qualification can be gained based on a proven track record of competency and accuracy in previous tasks, as rated by Amazon (Amazon.com 2014).

Figure 3-5: A sample of Mturk survey questions for the set “Diana Street”

The four street names used for each survey set are selected randomly with the prerequisite that they produce at least 20 matches per algorithm. Additionally, together they should cover the variations of the most frequently occurring syllable lengths (2, 3 & 4). The scores for each street pairing in each survey set are tabulated and averaged. The end product is a human annotated ranked/scored street name list which can be used as a benchmark for the lists produced by the other three algorithms. From there, the following two objectives are evaluated:

**Objective (1): precision and recall**

Precision/recall is evaluated via comparison of the original street name list produced by each algorithm to the entire collection benchmarked by Mturk participants. The concept of *all streets* and a *relevant street* is based on the following:
• **All streets:** The combined list of results of all algorithms (duplicates removed).

• **Relevant street:** An entry from *all streets* classified as “similar” based on human-annotation (average Mturk score above 1.0)

Precision of an algorithm would be the ratio of the number of relevant streets in its list to the size of its list itself. Recall would be the number of relevant streets in its list to the number of all relevant streets available. For example, if Mturk participants rated 35 streets as relevant (average score above 1.0), and algorithm A was responsible for 10 of those relevant entries from the 20 it produced, then:

- Precision for algorithm A = 10/20 = 0.50
- Recall for algorithm A = 10/35 = 0.29

**Objective (2): ranking of results**

The ranking quality of the new algorithm and Levenshtein Distance is determined via the comparison of the sequencing of Mturk results with both algorithms (Soundex cannot rank and is excluded from this test). For each of the four survey sets, the top scoring 20 results as determined by the Mturk participants are re-run thru both algorithms, thus producing two new lists of names with identical content but varying sequence of order. From there, the minimum switches/steps needed to reorganize these lists to mimic the order of the original Mturk list are calculated. A “step” is defined as the movement of a single street name exactly one row above or below its closest neighbor. For example, in Table 3-7, it takes a total of five switches to correctly reorder the days of the week (3 to move Wed up, 2 to move Sat down).

**Table 3-7: Example of step count for reorganizing days of the week**

<table>
<thead>
<tr>
<th>Original Order</th>
<th>move Wed 3 steps up</th>
<th>move Sat 2 steps down</th>
<th>Corrected order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon</td>
<td>Mon</td>
<td>Mon</td>
<td>Mon</td>
</tr>
<tr>
<td>Tues</td>
<td>Tues</td>
<td>Tues</td>
<td>Tues</td>
</tr>
<tr>
<td>Sat</td>
<td>Sat</td>
<td>Wed</td>
<td>Wed</td>
</tr>
<tr>
<td>Thurs</td>
<td>Thurs</td>
<td>Sat</td>
<td>Thurs</td>
</tr>
<tr>
<td>Fri</td>
<td>Fri</td>
<td>Thurs</td>
<td>Fri</td>
</tr>
<tr>
<td>Wed</td>
<td><strong>Wed</strong></td>
<td>Fri</td>
<td><strong>Sat</strong></td>
</tr>
<tr>
<td>Sun</td>
<td>Sun</td>
<td>Sun</td>
<td>Sun</td>
</tr>
</tbody>
</table>
The general performance of ranking is calculated based on Equation (4), whereby total steps is the number of steps needed to correct the order to that of Mturk, and max steps is the hypothetical steps needed to arrange the worst possibly arranged list of size \( n \) (see Equation (5)). For example, the ranking performance of Table 3-7 with a list size of seven, total steps of five and max steps of 21 is 23.8%.

\[
\text{Ranking Performance} = \left[ 1 - \left( \frac{\text{Total Steps}}{\text{Max Steps}} \right) \right] \times 100
\]

\[
\text{Max steps} = \frac{n(n - 1)}{2}
\]

3.3.1.1 Dealing with street suffixes

As explained previously in Section 3.2.1.5, confusion of street suffixes is considered to be less phonetic/character based, which would be a significant disadvantage for Soundex and Levenshtein. Soundex is almost incapable of providing proper suffix interpretation due to code length (Section 3.2.1.1) and Levenshtein is not capable of evaluating suffixes any differently from a normal name. For example, using Levenshtein Distance algorithm, the similarity of Brian with Bryan is 0.83 while Brian Street with Bryan Road is 0.50. This would severely handicap both algorithms throughout the entire evaluation process. However, the suffix should not be discarded due to the importance it plays in similarity perception (as will be proven in the Mturk results in Section 3.5.1.1) and the critical role it plays in our algorithm (quarter of entire score). Thus, two algorithm variations are employed for Soundex and Levenshtein Distance:

**Suffix is manually preset (Soundex and Levenshtein v1)**

The suffix handicap is removed by manually appending a symbolic suffix wherever needed for participant evaluation. This is applied on the Soundex and Levenshtein Distance v1 algorithm. Initially, the list of street names for the algorithms is generated purely without processing the suffix. Once the list is generated, the correct suffix for each street is manually appended in the survey sets. In the evaluation step for precision
and recall, Soundex and Levenshtein Distance are permitted to match purely based on street name. In the ranking step, the order is matched purely without suffix comparison (best case scenario). As a rule of thumb, whenever there was a disparity or uncertainty in matching due to suffix, the algorithms are always given the benefit of the doubt, and the shortest/best option is taken.

Suffix is processed (Levenshtein v2)
The street names and suffixes are combined into a single entity and processed as one word (e.g., “Bryan Street” is processed as “Bryanstreet”). This is applied to the Levenshtein Distance v2 algorithm. The method allows the natural handicap to occur in order to provide a fairer basis of comparison of each algorithm. The additional cost to the Levenshtein Distance algorithm would be based on the number of character insertions, deletion or substitutions necessary to change one suffix into another.

3.3.2 Application of algorithm on Greater Melbourne street name network
The new algorithm is then tested on a real street name network in order to uncover potential spatial attributes. The region chosen for testing the algorithm is Greater Melbourne, Australia. It has an area of 8,800 km² and is divided into 31 municipalities, consisting of metropolitan, urban and rural areas. It has a total population of 4.2 million people with a density of approximately 1500 citizens per km². The official street network data is obtained from the Public Service Mapping Agency (PSMA) of Australia in the shapefile format. The data contains the spatial information and the needed attribute information of street name, suffix type, and ID. The dataset is filtered for erroneous and overlapping data so that each unique street is assigned a single record entry and ID.

3.3.2.1 Handling street name particulars & exact duplicates
The Greater Melbourne street name database contains certain characteristics and naming styles as commonly found in most street networks that need to be addressed. Table 3-8 summarizes these characteristics and the action, or lack thereof, taken to handle them. For most issues, no specific action is necessary because it does not significantly impede nor skew the algorithm processing steps.
Table 3-8: Street name characteristics and plan of action

<table>
<thead>
<tr>
<th>Street name characteristic</th>
<th>Details</th>
<th>Action</th>
</tr>
</thead>
</table>
| Unnamed                    | • ~2000 unnamed streets (3.5%)
  • Yet to be named/left nameless
  • Mostly associated with utility lanes, back roads and rural paths | • These streets are considered trivial and were eliminated from database |
| Hyphen (“-“) Olinda-Monbulk, Orbell-Jones | • Hyphen usage is inconsistent across different street name sources | • Hyphen symbol removed |
| Initials                   | • Low occurrence (<0.02%) | • Ignore due to limited amounts |
| Numericals                 | • Low occurrence (<0.03%)
  • 95% located in a confined region used for farming and irrigation (Cocoroc locality) | • Ignore due to limited amounts |
| Generic prefixes The, Mt, De, El, La, Le, St, Little, Old, Glen, Lake | • Low occurrence (< 1%) | • Processed normally via algorithm |
| Multi-phrased              | • 1949 streets with multi-phrase (6.5%) | • Processed normally via algorithm
  Airport-Western Ring In    | • Entire metaphone phrase must match |

An unexpected high number of exact duplicates (streets with same name and suffix type) were observed that were sufficiently high to mask the results of the more important close but not perfect matches. As a perspective, over 200 street names have at minimum eight exact duplicates in Greater Melbourne. Since this research topic is aimed at uncovering subtle levels of similarity and exact duplicates are not hidden from authorities, steps were taken to limit their influence on the results. Thus, the algorithm is preset to not process a street with its exact duplicate more than once. Streets with the same name but differing suffix (e.g., Gordon Street and Gordon Lane) are permitted.

3.3.2.2 Conflict score acceptance threshold

The similarity rating of two streets is a continuous scale from zero to one and there is a point after which the level of association breaks down to an unacceptable level. A minimum threshold of 0.5 is set, based on the fundamental workings of the algorithm. To fall short of the 0.5 threshold, two of the four algorithm assessment methods would have to return an absolute reject or a combination of assessments would have to obtain extremely low scores. While street suffix assessment may be more subjective, the other three Metaphone, EndCode and n-grams are much more critical and any large compromise in these parameters would reveal a significant deviation from similarity.
3.3.2.3 Processing implementation

ArcGIS 10 is used as the primary platform to analyse and visualize the street data. The new algorithm is implemented as a separate stand-alone program and coded using Visual Studio. It has the ability to process any text-based database of words/names. The results produced are in .CSV (comma delimited) format and can be easily imported/exported between software. The spatial analysis is run using a combination of python scripts and the Visual Basic for Applications (VBA) code, which is an integrated component of ArcGIS.

Full street analysis is carried out by algorithmically cycling through every possible street pair from the entire dataset, each producing a unique score of similarity. Thus, each street would gain a score and match count that would represent its global conflict. The moving window is implemented by overlaying a grid on to the map, whereby each containing cell is of size 0.5x0.5km$^2$. A square-bounding window is created around each cell and all streets contained or intersecting the window are extracted and processed as a batch via the algorithm. The conflict score of all streets with each other is obtained, and then an overall score for the entire window is calculated as the sum of all individual street scores. This process is repeated with varying window sizes of 2.5km$^2$ (5x5 cells), 5.5km$^2$ (11x11 cells), 10.5km$^2$ (21x21 cells) and 15.5km$^2$ (31x31 cells). Data from the 2.5km$^2$ window (the most refined resolution) is further aggregated based on municipality boundaries. The municipality scores were further normalized for street density by dividing the score with the street count of that region.

3.4 Experiment results

Using some example streets from the dataset, Sections 3.4.1 and 3.4.2 highlight the performance of the new algorithm, Soundex and Levenshtein Distance in terms of scoring and ranking. Sections 3.4.3, 3.4.4 and 3.4.5 reveal the results of the algorithm applied to the entire Greater Melbourne street name dataset and the subsequent trends uncovered. Section 3.4.6 shows the possible phonetic trends of the street name network and Section 3.4.7 looks deeper into the effect of secondary duplicate clustering on all the results.

3.4.1 Performance results of algorithms

Table 3-9 shows the performance differences between Soundex, Levenshtein Distance v1, Levenshtein Distance v2, and the new algorithm based on the four street set surveys.
conducted in terms of precision and recall. Figure 3-6 provides a graph perspective of the overall averaged precision and recall results along with standard deviation (black dotted line). Table 3-10 is an example of the scored results for the *Craig Street* set.

Table 3-9: *Comparison of precision and recall values for all algorithms. Levenshtein Distance v2 represents results based on the inclusion of street suffix in processing.*

<table>
<thead>
<tr>
<th></th>
<th>Soundex</th>
<th>Levenshtein Distance v1</th>
<th>Levenshtein Distance v2</th>
<th>New algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>SET 1: Craig Street</td>
<td>0.13</td>
<td>0.23</td>
<td>0.50</td>
<td>0.45</td>
</tr>
<tr>
<td>SET 2: Barry Street</td>
<td>0.54</td>
<td>0.37</td>
<td>0.60</td>
<td>0.29</td>
</tr>
<tr>
<td>SET 3: Diana Street</td>
<td>0.37</td>
<td>0.60</td>
<td>0.55</td>
<td>0.36</td>
</tr>
<tr>
<td>SET 4: Ellison Street</td>
<td>0.83</td>
<td>0.26</td>
<td>0.80</td>
<td>0.42</td>
</tr>
<tr>
<td>AVERAGE</td>
<td><strong>0.47</strong></td>
<td><strong>0.37</strong></td>
<td><strong>0.61</strong></td>
<td><strong>0.38</strong></td>
</tr>
<tr>
<td>(Std Dev)</td>
<td>(0.29)</td>
<td>(0.17)</td>
<td>(0.13)</td>
<td>(0.07)</td>
</tr>
</tbody>
</table>

Figure 3-6: *Overall precision/recall results based on average of all four sets. The dotted lines represent the standard deviation.*
Table 3-10: Similarity results for “Craig Street” as produced by Mturk and all algorithms (Mturk score normalized over 1). Not all Soundex results are shown due to space constraints.

<table>
<thead>
<tr>
<th>Mturk</th>
<th>New algorithm</th>
<th>Levenshtein v1 (with suffix)</th>
<th>Levenshtein v2 (with suffix)</th>
<th>Soundex (no score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Craig Street</td>
<td>Craig Court</td>
<td>Craig Street</td>
<td>Craig Street</td>
<td>Carex</td>
</tr>
<tr>
<td>Craig Road</td>
<td>Craig Road</td>
<td>Craik Street</td>
<td>Braid Street</td>
<td>Cargo</td>
</tr>
<tr>
<td>Craigie Street</td>
<td>Craig Avenue</td>
<td>Craigie Street</td>
<td>Train Street</td>
<td>Carissa</td>
</tr>
<tr>
<td>Craig Avenue</td>
<td>Craig Close</td>
<td>Corang Street</td>
<td>Greig Street</td>
<td>Carriageway</td>
</tr>
<tr>
<td>Craig Drive</td>
<td>Craig Drive</td>
<td>Craine Street</td>
<td>Clais Street</td>
<td>Carrick</td>
</tr>
<tr>
<td>Craig Street</td>
<td>Creek Street</td>
<td>Craigton Street</td>
<td>Crane Street</td>
<td>Carrs</td>
</tr>
<tr>
<td>Greig Street</td>
<td>Crick Street</td>
<td>Duncraig Street</td>
<td>Haig Street</td>
<td>Ceres</td>
</tr>
<tr>
<td>Craig Court</td>
<td>Gurig Street</td>
<td>Craiglea Street</td>
<td>Cain Street</td>
<td>Cerise</td>
</tr>
<tr>
<td>Gregg Street</td>
<td>Greig Street</td>
<td>Train Street</td>
<td>Braim Street</td>
<td>Cherokee</td>
</tr>
<tr>
<td>Craig Close</td>
<td>Crook Street</td>
<td>Greig Street</td>
<td>Curtain Street</td>
<td>Chorus</td>
</tr>
<tr>
<td>Craigie Street</td>
<td>Corak Street</td>
<td>Chain Street</td>
<td>Clonaig Street</td>
<td>Chris</td>
</tr>
<tr>
<td>Greig Street</td>
<td>Carrick Street</td>
<td>Chaim Street</td>
<td>Corsair Street</td>
<td>Church</td>
</tr>
<tr>
<td>Haig Street</td>
<td>Gregg Street</td>
<td>Craft Street</td>
<td>Wraith Street</td>
<td>Cirque</td>
</tr>
<tr>
<td>Craig Road</td>
<td>Craik Road</td>
<td>Haig Street</td>
<td>Corris Street</td>
<td>Cirrus</td>
</tr>
<tr>
<td>Craigton Street</td>
<td>Kirk Street</td>
<td>Crana Street</td>
<td>Irving Street</td>
<td>Corak</td>
</tr>
<tr>
<td>Duncraig Street</td>
<td>Gourouck Street</td>
<td>Craie Street</td>
<td>Claire Street</td>
<td>Coris</td>
</tr>
<tr>
<td>Craine Street</td>
<td>Creek Street</td>
<td>Braid Street</td>
<td>Orange Street</td>
<td>Corris</td>
</tr>
<tr>
<td>Crane Street</td>
<td>Greig Court</td>
<td>Clair Street</td>
<td>Orange Street</td>
<td>Corrs</td>
</tr>
<tr>
<td>Crick Street</td>
<td>Creek Road</td>
<td>Cain Street</td>
<td>Carriageway</td>
<td>Courage</td>
</tr>
<tr>
<td>Gurig Street</td>
<td>Grigg Avenue</td>
<td>Braim Street</td>
<td>Cruise Street</td>
<td>(total of 70 results)</td>
</tr>
</tbody>
</table>

3.4.2 Ranking results

Table 3-11 reveals the ranking performance of Levenshtein Distance (v1 and v2) and the new algorithm as benchmarked to the Mturk survey results. A higher percentage value indicates sequencing closer to the “true” Mturk list (100% is equivalent to a perfect sequence match).

Table 3-11: Ranking performance of Levenshtein Distance and new algorithm.

<table>
<thead>
<tr>
<th></th>
<th>Levenshtein Distance v1 (total steps)</th>
<th>Average steps per entry</th>
<th>Levenshtein Distance v2 (total steps)</th>
<th>Average steps per entry</th>
<th>New algorithm (total steps)</th>
<th>Average steps per entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>SET 1</td>
<td>Craig Street</td>
<td>75.3% (47)</td>
<td>2.4</td>
<td>56.9% (82)</td>
<td>4.1</td>
<td>65.3% (66)</td>
</tr>
<tr>
<td>SET 2</td>
<td>Barry Street</td>
<td>50.5% (94)</td>
<td>4.7</td>
<td>60.5% (75)</td>
<td>3.8</td>
<td>53.7% (88)</td>
</tr>
<tr>
<td>SET 3</td>
<td>Diana Street</td>
<td>61.6% (73)</td>
<td>3.7</td>
<td>49.5% (96)</td>
<td>4.8</td>
<td>66.3% (64)</td>
</tr>
<tr>
<td>SET 4</td>
<td>Ellison Street</td>
<td>56.8% (82)</td>
<td>4.1</td>
<td>60.0% (76)</td>
<td>3.8</td>
<td>58.4% (79)</td>
</tr>
<tr>
<td>AVERAGE</td>
<td></td>
<td>61.1%</td>
<td>3.7</td>
<td>56.7%</td>
<td>4.1</td>
<td>61.0%</td>
</tr>
</tbody>
</table>
Table 3-12 is a sample set of the ranking for Diana Street before the process of re-sequencing. This reveals the relative starting positions of each street name and the extent of steps required to generate the correct order. The values in brackets represent the discrepancy between Mturk and current position.

Table 3-12: Sample ranking for Levenshtein Distance and the new algorithm as compared to the Mturk sequence.

<table>
<thead>
<tr>
<th>Mturk</th>
<th>Street name</th>
<th>Levenshtein v1</th>
<th>Levenshtein v2</th>
<th>New algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dianna Street</td>
<td>3   (2)</td>
<td>1   (0)</td>
<td>1   (0)</td>
</tr>
<tr>
<td>2</td>
<td>Diane Street</td>
<td>5   (3)</td>
<td>2   (0)</td>
<td>13  (11)</td>
</tr>
<tr>
<td>3</td>
<td>Dianne Street</td>
<td>11  (8)</td>
<td>7   (4)</td>
<td>17  (14)</td>
</tr>
<tr>
<td>4</td>
<td>Dinah Street</td>
<td>14  (10)</td>
<td>8   (4)</td>
<td>8   (4)</td>
</tr>
<tr>
<td>5</td>
<td>Daina Street</td>
<td>15  (10)</td>
<td>9   (4)</td>
<td>9   (4)</td>
</tr>
<tr>
<td>6</td>
<td>Diana Drive</td>
<td>1   (5)</td>
<td>17  (11)</td>
<td>5   (1)</td>
</tr>
<tr>
<td>7</td>
<td>Diana Court</td>
<td>2   (5)</td>
<td>18  (11)</td>
<td>2   (5)</td>
</tr>
<tr>
<td>8</td>
<td>Dina Street</td>
<td>6   (2)</td>
<td>3   (5)</td>
<td>3   (5)</td>
</tr>
<tr>
<td>9</td>
<td>Daina Court</td>
<td>16  (7)</td>
<td>20  (11)</td>
<td>12  (5)</td>
</tr>
<tr>
<td>10</td>
<td>Dana Street</td>
<td>7   (3)</td>
<td>4   (6)</td>
<td>4   (6)</td>
</tr>
<tr>
<td>11</td>
<td>Dianna Court</td>
<td>4   (7)</td>
<td>19  (8)</td>
<td>6   (5)</td>
</tr>
<tr>
<td>12</td>
<td>Giana Street</td>
<td>8   (4)</td>
<td>5   (7)</td>
<td>18  (6)</td>
</tr>
<tr>
<td>13</td>
<td>Dena Street</td>
<td>17  (4)</td>
<td>10  (3)</td>
<td>7   (6)</td>
</tr>
<tr>
<td>14</td>
<td>Dayan Street</td>
<td>20  (6)</td>
<td>15  (1)</td>
<td>20  (6)</td>
</tr>
<tr>
<td>15</td>
<td>Donna Street</td>
<td>18  (3)</td>
<td>11  (4)</td>
<td>10  (5)</td>
</tr>
<tr>
<td>16</td>
<td>Dina Retreat</td>
<td>9   (7)</td>
<td>16  (0)</td>
<td>11  (5)</td>
</tr>
<tr>
<td>17</td>
<td>Gianna Street</td>
<td>12  (5)</td>
<td>12  (5)</td>
<td>14  (3)</td>
</tr>
<tr>
<td>18</td>
<td>Kiana Street</td>
<td>10  (8)</td>
<td>6   (12)</td>
<td>15  (3)</td>
</tr>
<tr>
<td>19</td>
<td>Danae Street</td>
<td>19  (0)</td>
<td>13  (6)</td>
<td>19  (0)</td>
</tr>
<tr>
<td>20</td>
<td>Dianella Street</td>
<td>13  (7)</td>
<td>14  (6)</td>
<td>16  (4)</td>
</tr>
</tbody>
</table>

3.4.3 Overall Greater Melbourne street network conflict

Figure 3-7 reveals the global conflict levels of all streets in Greater Melbourne as matched with every other street in the region. The score of each street is the tabulation of all its matched conflict pairs, with each match assigned a value from zero to one. The dataset of streets have a mean score of 16.4 and standard deviation of 20.3. The maximum score obtained for a single street is 132.7. Visualization is done via groupings of streets based on ½ standard deviations, resulting in a total of nine classes. The colour range of green to red was used to reveal continuous increase in conflict score. Figure 3-8 reveals a zoomed-in perspective of the Melbourne city centre or CBD.
Figure 3-7: Visualization of full street name conflict levels of Melbourne

Figure 3-8: Street name conflict zoomed-in to Melbourne CBD
3.4.4 Distribution of matches per score threshold (graph)

Figure 3-9 shows the distribution of matches per street count for the entire Greater Melbourne region, using score thresholds from 0.5 to 0.9. This graph reveals the general conflict level of streets in Greater Melbourne and also the influence of the algorithm on the match distribution. A higher threshold would filter out low similarity conflicts, resulting in a smaller match count, and vice versa. Streets with zero matches are omitted due to space constraints.

![Distribution of Matches per Threshold](image)

**Figure 3-9: Distribution of matches for thresholds of 0.5, 0.7 and 0.9**

3.4.5 Conflicts by region (moving window results)

Figure 3-10 reveals the moving window results for the various window sizes. The extent of conflict is represented by the colour gradient, with green being the lowest conflict and red being the highest levels of conflict. The colour range is based on groups of ½ standard deviation classes, generating a total of eight classes. The blue lines represent municipality boundaries. Figure 3-11 represents the aggregation of the moving window data per municipality, using the finest window resolution (6x6 cells) and normalized for street density. The stronger red colour range represents a stronger conflict.
**Figure 3-10:** Moving window results. The main section represents 2.5km$^2$ window. The secondary section represents (from top to bottom) the 5.5km$^2$, 10.5km$^2$ and 15.5km$^2$ window size, respectively.

**Figure 3-11:** Aggregated results based on municipality with street density influence normalized
3.4.6 Phonetic trends

Table 3-13 and Figure 3-12 reveal the trends of the street name conflicts based on the length of their Metaphone codes. Phonetic algorithms are designed to represent the significant sound sequences of a word, thus its code length provides a rough representation of articulation length (i.e. longer codes are generally indicative of longer sounding street names). Table 3-13 depicts the five highest occurring phonetic codes in the entire Greater Melbourne street name network (in bold) along with their respective most frequent street names.

Table 3-13: The top occurring phonetics and associated street names for the entire Greater Melbourne street name network

<table>
<thead>
<tr>
<th>RANK (by frequency)</th>
<th>PHONETIC CODE LENGTH</th>
<th>PHONETIC CODE LENGTH</th>
<th>PHONETIC CODE LENGTH</th>
<th>PHONETIC CODE LENGTH</th>
<th>PHONETIC CODE LENGTH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 char</td>
<td>3 char</td>
<td>4 char</td>
<td>5 char</td>
<td>6 char</td>
</tr>
<tr>
<td>1</td>
<td>MR</td>
<td>KRN</td>
<td>KRTN</td>
<td>ALBRT</td>
<td>AKSFRT</td>
</tr>
<tr>
<td></td>
<td>Murra</td>
<td>Green</td>
<td>Gordon</td>
<td>Albert</td>
<td>Oxford</td>
</tr>
<tr>
<td></td>
<td>Mary</td>
<td>Crown</td>
<td>Garden</td>
<td>Alberta</td>
<td>Exford</td>
</tr>
<tr>
<td></td>
<td>Moore</td>
<td>Karen</td>
<td>Gardenia</td>
<td>Elberta</td>
<td>Axford</td>
</tr>
<tr>
<td></td>
<td>Marie</td>
<td>Curran</td>
<td>Curtin</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Moira</td>
<td>Kearney</td>
<td>Grattan</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>KR</td>
<td>BRN</td>
<td>FKTR</td>
<td>ALFR</td>
<td>ANTRSN</td>
</tr>
<tr>
<td></td>
<td>Gray</td>
<td>Byron</td>
<td>Victoria</td>
<td>Alfred</td>
<td>Anderson</td>
</tr>
<tr>
<td></td>
<td>Corre</td>
<td>Boroni</td>
<td>Victory</td>
<td>Alfreda</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Kerr</td>
<td>Brown</td>
<td>Victor</td>
<td>Ilford</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Grey</td>
<td>Byrne</td>
<td>Fitgtree</td>
<td>Alleford</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Quar</td>
<td>Barnia</td>
<td>-</td>
<td>Eleford</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>TN</td>
<td>KRL</td>
<td>MRTN</td>
<td>ALZB0</td>
<td>STRLNK</td>
</tr>
<tr>
<td></td>
<td>Dean</td>
<td>Coral</td>
<td>Martin</td>
<td>Elizabeth</td>
<td>Stirling</td>
</tr>
<tr>
<td></td>
<td>Diann</td>
<td>Carroll</td>
<td>Merton</td>
<td>-</td>
<td>Starling</td>
</tr>
<tr>
<td></td>
<td>Dawn</td>
<td>Corella</td>
<td>Morton</td>
<td>-</td>
<td>Sterling</td>
</tr>
<tr>
<td></td>
<td>Dion</td>
<td>Crawle</td>
<td>Moreton</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Don</td>
<td>Curlew</td>
<td>Meridian</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>KL</td>
<td>MRN</td>
<td>KRNT</td>
<td>MRKRT</td>
<td>WLNKTN</td>
</tr>
<tr>
<td></td>
<td>Cole</td>
<td>Marion</td>
<td>Grant</td>
<td>Margaret</td>
<td>Wellington</td>
</tr>
<tr>
<td></td>
<td>Kelly</td>
<td>Marina</td>
<td>Garnet</td>
<td>Margareta</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Clay</td>
<td>Marne</td>
<td>Grand</td>
<td>Margaritta</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Koala</td>
<td>Maureen</td>
<td>Kurand</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Kylie</td>
<td>Marine</td>
<td>Coronet</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>JN</td>
<td>BRT</td>
<td>BRTN</td>
<td>FRNSS</td>
<td>FLNTRS</td>
</tr>
<tr>
<td></td>
<td>John</td>
<td>Brett</td>
<td>Barton</td>
<td>Francis</td>
<td>Flinders</td>
</tr>
<tr>
<td></td>
<td>Joan</td>
<td>Barrett</td>
<td>Burton</td>
<td>Frances</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Geno</td>
<td>Baird</td>
<td>Brighton</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Jean</td>
<td>Bright</td>
<td>Baradine</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>June</td>
<td>Bird</td>
<td>Brought</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 3-12 shows various attributes associated with the different phonetic code lengths. The first y-axis reveals the average number of conflicts that occur for a street based on its phonetic code length. The second y-axis reveals the quality of these conflicts based on average score
obtained. For example, streets with phonetic code length of two like MR (Murra, Mary, Moore) or KR (Gray, Corre, Kerr) have on average 70 conflicts each, of which most are of low quality matches (score less than 0.6).

![Figure 3-12: Conflict count based on phonetic code length](image)

3.4.7 Secondary duplicate clustering effect

As explained in Section 4.2.1, exact street duplicates were prohibited from matching with each other more than once in the algorithm. However, this did not eliminate the possibility of a street closely (but not perfectly) matching a large cluster of duplicates in a one-to-many scenario. For example, Garden Street could potentially produce twenty matches of similarity, of which fifteen are all Gordon Street but in various suburbs. This significantly raises its overall conflict score. This secondary duplicate clustering effect in Greater Melbourne is shown in Figure 3-13. The x-axis represents the 30,000 streets with the highest conflict score (from left to right). The y-axis reveals the percentage of their score that was influenced by this secondary clustering issue. Streets with conflict score of zero are not displayed. The red line represents an averaged trend line of this effect (based on a normalized 500-record moving average). On average, the top five thousand streets can attribute ~50% of their conflict score to the effect secondary duplicate clustering. This means that most of the street name conflict exiting in the entire region is not only a result of the many exact duplicates matching with themselves (e.g.,
Moore Street occurs 13 times), but also with their similar counterparts in large quantities (e.g., a single Mohr Street matches to a high degree of similarity with each of the 13 Moore Streets).

![Figure 3-13: Analysis of secondary duplicate clustering on results](image)

### 3.5 Discussion of algorithm performance and spatial trends

Section 3.5.1 discusses the algorithm in detail and its performance differences with Soundex and Levenshtein Distance, as benchmarked with Mturk results. Section 3.5.2 reveals the outcome of utilizing this algorithm in tandem with spatial techniques on the Greater Melbourne street network. The trends and characteristics of the dataset and other observations made are evaluated in Section 3.5.3.

#### 3.5.1 Overview of algorithm performance

Table 3-9 and Table 3-11 represent the general performance of all algorithms as compared to a human-annotated set (Mturk survey). Overall, the new algorithm provides an increase performance of roughly 15–35% in precision and 10-25% in recall when compared with Soundex and both Levenshtein Distance versions. The standard deviation for precision and recall is 16% and 8% respectively. As a relative comparison, when Snae (2007) tested Metaphone directly with Soundex, there was a minor 5% reduction in recall as collateral for
gaining an increase of 12% in precision. While their objective was slightly different (uncovering name equivalency as opposed to name similarity), the comparison of their results versus this new algorithm does show promise. Levenshtein Distance v2 (with suffix processed) had an overall lower performance of 17% for both precision and recall when compared to its alternative Levenshtein version.

Looking at each set individually, the new algorithm outperforms the other algorithms in all areas except for the Ellison Street set (precision discrepancy of 8% with Soundex, recall discrepancy of 3% with Levenshtein Distance v1) and Diana Street set (recall discrepancy of 13% with Soundex). These two shortcomings can be attributed to both street sets containing a uniquely high amount of phonetically non-equivalent results (e.g., Ellison with Eddison, Diana with Gianna) that were deemed relevant by the Mturk participants (score above 1.0). In these cases, the Levenshtein Distance method is more successful, since it is tolerant of minor character changes, while the Metaphone component in the new algorithm would reject this match entirely. Meanwhile, Levenshtein Distance v2 (suffix processed) came in last place for all areas except for the Craig and Barry sets (coming in second-last behind Soundex in precision values). This shows that, as expected, the value assigned to a suffix is more than its sequencing of character, spelling or sound.

The ranking aspect is essential because it gauges how well filtered the results are and which entries are suppressed/highlighted. The overall results are less distinct than expected (Table 3-11). Levenshtein Distance v1 was the overall best performer by a narrow margin (0.1% better than the new algorithm). However, the comparative performance of each algorithm was extremely varied per street set revealing the difficulty of standardizing similarity methods when it comes to detailed comparisons. The new algorithm slightly edges Levenshtein Distance v1 in switch count in all sets except for the Craig Street set. This can be attributed to the amount of suffix variations available for matching which, as discussed in Section 3.3.1.2, allowing Levenshtein to benefit from the provided advantage (almost half of all Craig Street entries in the top 20 are of non-matching suffix).

Table 3-12 reveals a sample ranking of Diana Street as compared to Mturk results along with the disparity of position to the “true” Mturk rank. It can be seen that the major ranking disparity for Levenshtein Distance v2 occurs mid-table (where suffix starts to vary) and for the new algorithm it occurs upper-table (phonetics slightly varies, suffix is still the same). This
confirms that participants are choosing to prioritize street suffix above minor phonetic changes. This issue is further discussed in the following section. In general, this is where future work can focus on improving the weighting of each component to better mimic human perception of confusion.

3.5.1.1 Analysis of results

Looking into more detail, there are some general traits and patterns that can be observed from how Mturk participants are making decisions regarding similarity. At quick glance, the new algorithm eliminates major outliers as compared to Soundex (e.g., Soundex matches Craig with Crossway, Crusoe and Corang). However, more specific trends can be uncovered when characteristics of the original street name are compared to all its associated relevant matches (Scored above 1.0 by Mturk participants). Table 3-14 highlights some of these significant trends that will be discussed, which include beginning and trailing phonetics of a street name, the influence of suffix, and consistency of syllable count.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Rate of occurrence</th>
<th>Highest position of occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(All relevant results)</td>
<td>(Top 10) (Average over 4 street sets)</td>
</tr>
<tr>
<td>Matching first character</td>
<td>97.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Matching EndCode</td>
<td>86.0%</td>
<td>90.0%</td>
</tr>
<tr>
<td>Matching street suffix</td>
<td>73.0%</td>
<td>70.0%</td>
</tr>
<tr>
<td>Matching syllable count</td>
<td>72.0%</td>
<td>80.0%</td>
</tr>
</tbody>
</table>

In terms of first character, 97% of all relevant results have a first character that matched the original street name in terms of a vowel-vowel or consonant-consonant grouping (e.g., Diana, Gianna, Kiana all begin with consonants while Ellison, Allison, Elysian all begin with vowels). This further reinforces the widely accepted notion of the role of the first character in phonetically defining the uniqueness of a name. The more pertinent issue would be to evaluate the designed EndCode component, which instead parameterizes the trailing sounds of names. In this case, over 86% of the relevant results are in compliance with the EndCode vowel grouping. From those not aligned to this grouping, only 10% are ranked in the top ten. This shows that the trailing sound is highly valued in similarity, almost as much as the initial sound of a name. Some examples of outliers not in compliance with EndCode are Craig Street
(EndCode = 0) with Craigie Street (EndCode = 3), and Diana Street (EndCode = 1) with Diane Street (EndCode = 0). There is also potential future work in dealing with the trailing letter of a street name that can sometimes merge into the starting letter of a suffix (e.g. the letter [s] in Craigs Street), hence almost masking the phonetics of the trailing sound.

The street suffix is also influential in ranking of the Mturk survey. When given a choice, participants frequently opted to overlook minor phonetic differences when there was opportunity to match a street suffix (e.g. when evaluating the Ellison Street set, Elletson Street is ranked above Allison Road. There also seems to be a certain order of preference toward what can be confused with the suffix Street. Given the same street name but with variation of suffix, participants chose to rank according to the results shown in Table 3-15. Overall, the suffix road is rated to be the most related, court to be the least, and lane/drive/avenue to be intermediate. This pattern does adhere to the theory of frequency of occurrence to an extent, as assumed in the new algorithm. However, it is a simple yet telling confirmation that not all street suffixes are considered equal in the mind of users. The Levenshtein Distance v2 algorithm is handicapped in this area because the cost of changing all characters of one suffix to another is so high. This almost eliminates all possibility that a street name pair with non-matching suffixes can obtain a high rank.

### Table 3-15: Ranking of street suffix based on user perception (Mturk survey) for duplicate street names with varying suffix

<table>
<thead>
<tr>
<th>Street name set</th>
<th>Ranking method</th>
<th>Ranking results (ordered high to low)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Craig</td>
<td>Mturk survey</td>
<td>road, avenue/drive, court</td>
</tr>
<tr>
<td>Ellison</td>
<td>(human-annotated)</td>
<td>road, lane, drive, avenue</td>
</tr>
<tr>
<td>Barry</td>
<td></td>
<td>lane, court</td>
</tr>
<tr>
<td>Diana</td>
<td></td>
<td>drive, court</td>
</tr>
<tr>
<td>ALL STREETS</td>
<td>New algorithm</td>
<td>court, road, avenue, drive, lane</td>
</tr>
<tr>
<td></td>
<td>(by frequency of occurrence)</td>
<td></td>
</tr>
</tbody>
</table>

While syllable count is not a part of the new algorithm, there are significant trends to show its relevance in similarity. More than two thirds of all relevant results have the same syllable count as the original street. An example of this is shown in Table 3-12 based on Diana Street. Also, from the collection that did differ, 93% are either one or two syllables lower than the original name they are being compared to. For example, the original street name Diana (three
syllables) is associated with Diane (two syllables, ranked second) and Dinah (two syllables, ranked fourth). Adding to that, most (58%) of these syllable discrepancies are for the street set Diana which has a flexible interpretation of pronunciation/syllable count (*dai-nah OR di-an-ah*).

**3.5.1.2 Experiment integrity**

There are many other factors that could skew or affect the produced results. The suffix advantage (as discussed in section 3.3.1.2), which always provides a best-case-scenario option for Soundex and Levenshtein Distance v1, does create a significant bias, especially during processing. The more variations a street name has (e.g., *Craig Road, Craig Lane, Craig Avenue*), the higher the potential of the new algorithm being outclassed in ranking. Additionally, the rating scale used for the survey and its associated definition phrasing will indefinitely influence the overall outcome. For example, the cut-off score of relevancy at score 1.0 is the difference between “low probability of confusion” and “unlikely probability of confusion” which can be open to personal interpretation by each participant. This will all play a role when deciding what cut-off score to use for categorizing a “relevant” street name.

From the survey point of view, there can be some possible discrepancies caused by its design and how it is received/interpreted by participants. Participants were required to read the text, mentally imagine this verbalized scenario and infer an outcome based on a projected hypothetical situation, as opposed to actually experiencing the confusion themselves. This can make the results akin to a second hand response that might not necessarily be a perfect reflection of how conflicts play out. However, for now, this survey method still provides an acceptable gauge of confusion based on quality and feasibility of experiment.

Also, pronunciation competency is not a certainty for each participant, especially given his or her anonymous education level and cultural background. Participants were required to have masters level qualification (assured level of competency and literacy by Amazon), but this does not necessitate they would know how to interpret all the culturally varied street names involved. In fact, some participants provided feedback regarding accents, stating “...listening to a 911 call from somewhere in say Kentucky would sound a lot different from a call coming from New York City...”. This showed that, to some degree, some participants made interpretation based on accents.
Finally, time lapse and memory capacity can heavily influence how street names that were heard are remembered. For example, it would be much harder to mistake a John Street with John Lane during a discussion itself, but confusion could potentially occur if a user had to recall said street name among a list of others a few days later.

### 3.5.2 Overview of Greater Melbourne street name network

The algorithm is then applied throughout the entire street name network in an attempt to find patterns of conflicts. Overall there are 40,277 (69.4%) unique streets in Greater Melbourne in terms of name and suffix type. Figure 3-9 shows the overall score distribution for the dataset, as a function of successful matches per street. All thresholds reveal a long-tailed negative gradient, revealing a limited number of streets (15%) are responsible for a major part (~50%) of the street similarity conflicts. Regardless of threshold levels, the majority of streets in the dataset attained five matches or less. There is a marked reduction in streets with 15 matches or more, never exceeding 1000 streets (less than 2% of total dataset).

There are large variations in this distribution between the 0.9, 0.8 and 0.7 score thresholds. However, the subsequent thresholds of 0.7, 0.6 and 0.5 have very similar patterns. For streets obtaining 40 matches and above, the differentiation of scores become less apparent and threshold trends begin to merge. This can signify that the algorithm has the ability to effectively detect subtle differences in high scoring matches (0.7 and above) but for scores lower than that, the similarities start to become more vague. Overall, it is a positive improvement, revealing the variations in high scoring streets and the ability of the algorithm to pick up these subtleties, as opposed to a homogenous collection of acceptable matches as produced by Soundex. The majority of these low match-count but frequently occurring conflicts can be attributed to a few influential factors. Among these were: a) exact street name duplicates, b) city grid replication phenomenon, c) English centric naming, and d) two-character phonetics (discussed in Section 6.2.2).

Exact street name duplicates cause a bigger detriment than expected in terms of the secondary duplicate clustering effect. Figure 3-13 reveals that for the top 5000 streets, there is a clear strong influence of these duplicates, from a range of 20% up to 90%. Based on the trend line, the average influence steadily declines from 50% to 20% for streets ranked 1 to 25,000. This correlation rapidly starts to break down for lower ranked streets. In certain cases, the entire conflict score is attributed to this phenomenon, as can be seen in the series spikes
reaching the top of the $y$-axis. On average, 50% of the score of the top 5% conflicted streets can be associated to this issue. Thus, duplicates not only clash amongst each other, but their amounts are large enough to have significant knock-off effects on all their associated close matching neighbours throughout the region. This means eliminating exact duplicates has a subsidiary benefit of stabilizing conflict levels of their similar sounding counterparts.

There is also an issue of intentional replication of city street names, whereby developers opt to recreate a sense of familiarity based on names that people are comfortably associated with. The 13 streets that make up the rectangular city grid which include Bourke, Collins and Elizabeth have an average of 32.6 exact duplicates, putting them in the top 25% scorers of conflicted streets. Occasionally the entire city grid name collection is adopted in a neighbourhood (Banham 2013) as occurring in suburbs of Cheltenham, Bulleen, Coburg, Essendon and others.

In addition, there is an additional socio-political perspective to the street naming conflicts in Greater Melbourne (Table 3-13). There is clear dominance of a selection of generic English-centric names that were “imported” from England in early years and continuously recycled. Names such as Victoria Street (named after Queen Victoria) and Park Lane represented people and label styles not unlike the streets in England. The street names were initially motivated by acts of tribute to homeland icons and also to create a more familiar environment for the colonist (Greig 1941). However, the influence from a single cultural-lingual source creates phonetic homogeneity such as can be seen in clashes of Burton-Barton, Merton-Morton and the other top occurring names in Table 3-13. Unsurprisingly, results showed that English names rarely clashed with the aborigine-originated names, showing that cultural diversification of street names naturally reduce street confusion. Arguably, while Greater Melbourne has been undergoing a high rate of immigration of varied ethnicities/nationalities for the few past decades (Jupp 2007), it has had limited influence on street naming. For example the Asian dominated suburb of Boxhill and the Greek dominated suburb of Oakleigh do not have a single culturally referenced street name.

**3.5.2.1 Spatial distribution of conflicts**

In terms of spatial distribution, Moran’s I test was used to determine the spatial autocorrelation of street scores. This method assumes a normal distribution of the dataset and evaluates the probability that the outcome is either clustered or randomly set. The street
scores produced a Z-score of 66.73 indicating strong influences of clustering (less than 1% likelihood the clustered patterns are a result of random chance).

As can be seen from Figure 3-10 and Figure 3-11, the strongest area of conflict is located at the city centre with slightly lower intensity patches scattered in the surrounding suburbs. There is also a significant conflict patch in the south east tail of Melbourne (Frankston). The high conflicts in the city spill out to the neighbouring areas of Port Phillip, Yarra North and Boorondara. On average, the metropolitan streets had 29.87 matches, urban areas had 28.45 matches and rural areas had 25.76 matches per street, revealing a gradual tapering effect when moving away from the city. From Figure 3-11, even when the scores are normalized for density, it still shows the central regions dominate, showing the amount of streets is not necessarily influencing the results.

This overall result can be attributed to a number of causes. In general the central municipality has more administrative neighbours, creating more opportunity for clashes with its surroundings. Also the centralized regions were developed earlier, prior to awareness of duplicate issues. Once conflicting streets were established in the community, it was harder to gain an imperative for change. Finally, there is a tendency to replicate city street names as previously discussed in Section 3.5.2.

3.5.2.2 Phonetic patterns of conflicts
There are certain spatial-phonetic trends that emerge from the analysis of the street name network. As expected, except for single (1) character codes, the average number of conflicts decreases with increasing code length (Figure 3-12). This can be attributed to the lessening number of long street names available to create a conflict. However, the quality of the matches becomes higher. Thus, longer sounding street names have less pairs of conflict with each other, but those that did match were of higher quality match.

Table 3-13 depicts the five most frequently occurring phonetic codes in ranked order. Phonetics with less than two and more than seven characters were of too small quantities and considered trivial. Listed below each code are the top occurring street names associated with those codes (based on frequency count, not algorithm scoring). The two-character phonetics reveals a large collection of very varied conflict quality matches. While the full-implemented algorithm is able to make a better assessment of the results, the fact that they fulfil the
Metaphone criteria and are of high occurrence reveals the general low quality matches that they create. They also contain less effective consonant sounds (as inferred by the phonetics), therefore the remaining vowel sounds play a more significant role in the overall name pronunciation. Excluding the implemented EndCode system, the remaining vowels are not quantified. This can be reflected in the varying names Murra, Mary, Moore and Marie which all fall under the same phonetic code “MR”. In this case the developed algorithm is limited to observing name spelling variations via n-grams which still generates rankings relatively well but does not inherently address the real issue of understanding all vowels.

Long street names are defined as streets whose phonetic code lengths are the outlying 5% as assessed from a normal distribution. Phonetic codes of lengths 9 and above fulfil this criterion. While they produce a smaller collection of possible matches, the resultant street pairs are of high similarity. Thus longer sounding names, if not handled properly, can be even more dangerous because it creates a sense of false security in its length. There is a need to find a balance of name length in terms of ease of usage and phonetic uniqueness.

3.5.3 Further algorithm evaluation (vowels and street suffix)

Continuing on our argument regarding vowels in Section 6.1.1, some direct positives can be seen in the EndCode implementation. This is evident in its ability to distinguish the top occurring three-character phonetics of BRN in the Greater Melbourne street name network (Table 3-13) via the distinguishing of Byron (Endcode 0), Boroni (Endcode 3) and Barnia (Endcode 123) that previously could not be done in Soundex.

Overall the EndCode distribution ratio from code zero to code six is 35% (consonant or undetermined), 52% (“ah”), 3% (“ay”), 6% (“ee”), 3% (“ai”), 1% (“ow”) and 0.2% (“oo”) respectively, with consideration that a name can be in more than one group. These results show trailing vowel types are quite unevenly distributed and the extent of filtering is dependent on the popularity of the vowel group. The potential next step is to classify vowels in all character positions, and not only the trailing ones.

Street suffixes managed to create an explicit evaluation of street suffix, which is previously ignored in VICNAMES’ Soundex system. Hence more common suffixes like road, street and avenue would generate a stronger association level amongst each other compared to hill, pass and vista. However, the street suffixes could benefit from a better perceptive interpretation,
as opposed to being purely evaluated by usage count. While the highly used street and road are valid situations of mix-up, the similarity of some suffixes are less indicative of their similarity in frequency count, like Lane (948 occurrences) and Crescent (933 occurrences). Adding to this, the frequency distribution of suffixes is extremely heavy tailed, so that those not in the top 14 (16% of suffix types) cannot effect the final suffix score more than 1%. The practical outcome of this method is akin to a split class system that leads to only three possible outcomes: high/high, low/low or high/low match. Thus the scoring range becomes less continuous.

Additionally, there is a contentious assumption that confusing a street suffix pair is symmetrical or unidirectional, especially between a unique and common suffix. It may be feasible that a person forgets esplanade suffix and inadvertently substitutes in with street suffix, but for that to occur inversely is unlikely. The only exception would be if a user is explicitly provided with a list of alternate options or receives other external stimuli.
CHAPTER 4

CONCLUSION

This thesis aimed to investigate specific problems associated with reconciling place names communicated by a user with a dataset of existing place names in spatial databases. The research targeted the issue of confusion from place names that sound similar (street name phonetic similarity) and the use of circumstantial place names that are temporally dynamic as an alternative to the official name (place-event substitution).

4.1 Overall summary of research topics

On topic 1 regarding detecting event names that take on the role of place names, it has been found that Twitter can be used to observe the evolution in the place name terminology among its users. The approach presented in this work is based on the understanding that Twitter, as a social media platform, can be used as a representation of a community and its communicative methods. Through the use of the Twitter Streaming API, burst detection modelling and terminology extraction methods, the changes of place name choices were observed.

The Twitter discussion peaks do closely match the actual kick-off times of events (with a variance of hours) and speediness of detection is highly dependent on the type of event in question. Events that are result orientated, having a wider following, and are broadcasted live tend to generate discussion spikes closer to and during the event occurrence itself. This assumption could possibly be extended to popular entertainment award ceremonies relating to film, television and music (e.g., Academy Awards, Music Video Awards, Grammys, Emmys). It could also be generalized for team sports with leagues that generate sufficient following (e.g., football, basketball, rugby) especially if the teams in question possess a consistent Twitter fan base. In all, burst detection methods are well suited for multi-purpose venues that are designed as an environment for human-centric activities.

However, the overall results also confirms the volatile state of social media as a data source. It is hard to find an exact fit of statistical parameters for events because they are heavily influenced by the unpredictability of public interest levels. These interest levels can in turn be influenced by media coverage, event marketing and peer-to-peer discussions. Parameters can potentially be streamlined based on common fixed characteristics shared by events. So, for
example, concert-based events might share similar patterns of discussions that peak 24 hours before and after their occurrence, as shown in the experiments. Therefore, there might be justification to categorize place-event substitution by its characteristics like event type, event popularity and venue capacity in order to better model the discussions.

In terms of tweet rate patterns, the difference of tweets versus an interpersonal discussion results in low discussion rates for arenas and theatres if no event is imminent. In these situations Twitter becomes a poorer representation of layman discussions, resulting in place names being hard to detect and model during event downtime. In terms of Twitter streaming, the data collection method of keyword search is susceptible to ambiguity and has to be augmented with strong semantic or context filtering (natural language processing) to eliminate false positives. For this experiment, the combination of selecting unique place names and semi-automated training of context was sufficient. The event names were statistically easy to distinguish from among the other commonly used words. Expanding this scope still face challenge, especially in resolving the ambiguity of language phrasing and continually evolving aliases.

Work on topic 2 attempts to better understand the issue of street name confusion due to perceptual and spatial similarities, and improve upon the current Soundex algorithm. The major achievement is the improved performance in precision (15-35%) and recall (10-25%) compared to the other algorithms, as benchmarked with a hand-annotated test. This not only helps users more efficiently resolve street name uncertainty, but also provides a better gauge of similarity clustering patterns by region. This holistic street name evaluation is achieved by combining diverse evaluation methods into a unified scoring system. This included the integration of existing phonetic algorithms (Metaphone), self-developed component for parameterizing trailing vowel sounds (EndCode), use of n-grams for generic spelling closeness and the parameterization of the street suffix effect. The ranking performance was less concrete, with fluctuating results depending on street set evaluated. The new algorithm came in second place by a narrow margin (0.1%) for overall ranking quality. Even though all of the components of the new algorithm are well justified, the level of influence each component should have on the overall score is less certain. This highlights the difficulty in settings weights for the more refined perceptual differences in similarity.
The implementation of the algorithm on a real street name network further reveals the possibilities of understanding real-life street name conflict at a similarity and spatial perspective. The Greater Melbourne street name network revealed a highly conflicted metropolitan area that slowly tapers when extending outwards from the city to the suburban and rural areas. The assumption that this is due to the city containing more streets per unit area is proven false since these patterns were unchanged even when the score was normalized for street density. A small number of street names are responsible for the most conflict, especially when considering the English-centric duplicates. Additionally, duplicates play a bigger role than anticipated, both directly and indirectly and should be taken into consideration due to their knock-off effect on other close sounding matches. This new algorithm also helped to reveal the potential phonetic groupings and types of names to avoid per region. Government policies seem to be effective on streets for newer areas of development. Based on region type/distance policies, the city centre and its surroundings seem to fail in meeting this criterion. The relevant authorities have to decide on course of action, be it to change or keep existing streets, both with its sets of challenges.

4.2 Contributions and challenges towards more adaptive spatial databases
As discussed in Section 2.1 and Section 3.1, place name usage cannot be standardized in society, with vast variations and interpretations on how a location can be labelled or pronounced. Current gazetteers and spatial databases (see Section 2.1.1 – 2.1.3) are inherently static and have limited scope of knowledge of these variations beyond official documentation (e.g. gazetteers) and what is provided by a confined group of administrators or active data contributors (e.g. Geonames). Hence, there is an importance to reactively “know” the latest coined place names or dealing with multiple pronunciations to the said place name. Industries requiring accurate geospatial knowledge based on human input will continue to grow, which includes the reconciliation of place name delivered via speech/text into a location (e.g. location based services, mapping projects, tracking & navigation, etc). Thus, there is merit in creating more adaptive infrastructure that can better deal with uncertainty or ambiguity of human-machine communication.

There has been related work in bridging this communication gap, especially with large databases. More specifically, research and development has focused on processing oral and written user input. Voice recognition is found to be one of the most efficient methods of information transfer (Cohen and Oviatt 1995) and is predicted to be the critical form of
communication for future interactions with machines. Its advantages over a graphical user interface (GUI) like touch screen or keypad is its speed and flexibility of input (understanding free-form descriptions). There is also the inverse of training machines to properly pronounce names (text-to-speech) like reading a telephone book (Vitale 1991). The major struggles in this area of voice/oral communication have been on deciphering phonemes, distinguishing syllables and words in a sentence and understanding context so that spoken word can be matched/reconciled with known entities (objects/places/people). These challenges are linked to many practical daily-life scenarios such as identifying a person by name, finding a nearby restaurant or looking up an item in a shopping catalogue using a hand-held device or computer.

This research thesis contributes to this field by augmenting additional place name knowledge to geospatial databases. The additional knowledge is both in dynamic alternative place names (topic 1) or variations/errors in pronunciation of existing place names (topic 2). This builds on the understanding that humans have a vast of vocabulary that often is not standardized, and the choice of words/phrases they choose in order to describe a similar object can vary (Furnas et al. 1983, 1987). Research has shown that when random pairs of people where asked to name a single object, only 10-20% of the results were similar (Furnas et al. 1987). This problem is believed to cross-over to alternative place names and pronunciation among society. Hence there is a continual need for databases to be equipped with the technology to adaptively acquire new knowledge of human phrasing/wording in order to increase the success rate of machine interpretation.

While there are many types of alternative place name categories (as discussed in Section 2.1), the theme of event names being used as a place name (topic 1) has been given less attention in the research community, especially when extracting and localizing a phrase/term to a geographic location that already has an official place name. Event names are also the most evasive due to its short life span. There is strong relationship between event, place name usage and time (Jin et al. 2007), of which all three can be considered to mutually support each other. This co-relation has been shown in event detection in newspaper articles too (Jin et al. 2007). This research further strengthens this notion that place name and event name can go hand-in-hand and in the right circumstances can actually precede the other for a specific time period.
The concept of mining new knowledge and events from social media is not new, but has had limited forays in monitoring and tracking the change of place name usage over a time period. The nature of social media and Twitter makes it an appropriate choice of tracking real-time changes in cultural discussions, which includes the latest spatial locators used. At its current form, the designed system is able to monitor place name changes for amphitheatres, stadiums and sites of mass events. However it would be less applicable for places with events of low participation, since there is a dependence on a minimum amount of tweet discussions needed.

Besides that, the place name chosen might correct, but the method of pronunciation might vary, as researched in topic 2. Since the current default algorithm for databases including that of VICNAMES is still Soundex (which is widely accepted as poor quality), there is opportunity for improvement. Building on previous research of string matching and other phonetic algorithms, an improved algorithm is developed which is specifically designed to work on street name conflict.

Street names are not the only cause of location conflict/confusion in an urban environment, but it is seen as the most applicable and prominent, as seen from frequent real-life cases of tragedies as a result of misdirection (Coveyduck et al. 2004; Toronto Star, 2006; Ciccone 2012; Flowers 2012). However, since the developed algorithm works on any delimited text file, there is flexibility to be integrated to any place name database, with the minor removal of the street category component of the algorithm.

In its current form the algorithm be directly integrated into current address databases of emergency responder call centres. In theory, a caller would quote a street name which would then be heard by the responder and keyed-in into the system. The new algorithm would act as a mediator between the input text and list down (in ranked order), the most probable matches. Therefore any mispronunciation, misinterpretation or misspelling of a street name (from the side of the caller and/or the receiver) can be compensated by the algorithm. This method is considered an assisted human-machine interaction, since the place name information goes thru a trained mediator before the actual machine algorithm. This algorithm can potentially be augmented with voice recognition technology in order to bypass a human receiver. The compounded error of using two automated systems (new algorithm with voice recognition) might be significant, so caution must be taken with the current state of
technology. Even longer-term prospects are possibility of augmentation with natural language technology (e.g., handling dialog, stuttering, pauses, utterance, breaking down of sentences).

Overall these two topics allowed an in-depth study of phenomena where named entity resolution based on well known and authoritative place names would fail. Apart from these two topics, there are many other related phenomena that can also trigger these problems. Places can have unofficial but widely used nicknames like The Big Apple (New York City), Hells Kitchen (neighbourhood of Clinton, Manhattan) and The Razor (Gillette Stadium, Massachusetts). Official place names can also undergo a public reinterpretation like Micky Ds for Australian McDonalds and TriBeCa, which is vernacular for “Triangle Below Canal Street” in New York City. Additionally, mispronunciation of place names can occur due to differences in language, culture and slang of the user, which could radically change the interpretation of what is communicated. For example the locality of Prahran, Melbourne is commonly mispronounced as “Prah- run” by non-locals, as opposed to the correct “Pran”. A group of places can also hold a generic label along with their own official individual place names (e.g., the “Caribbean islands” consisting of many small official islands). There can also be duplicate place names that can only be distinguished by their higher granularity grouping (e.g., a Park Lane street in each suburb). All these cases represent the diversity of scenarios that require improvements to the current spatial database systems.

4.3 Future work

The future work for the place-event detection system can look at observing different types of place name usage, especially in the areas of duration and co-occurrence. Certain changes have a more stretched duration and can span months to years. Some primary examples are official name changes, dual-name places and nicknames. Official name changes occur when authorities replace an existing place name with a new one, usually motivated by political or commemorative reasons, such as the change of the London Clock Tower/Big Ben to the Elizabeth Tower in 2012. Dual-name place names are usually a result of a naming compromise between a native and colonist group, whereby both names are official and equally acknowledged (Uluru/Ayers Rock in Western Australia, Denali/Mt McKinley in Alaska). Nicknames, as discussed previously, are a result of a community/fan base manifesting emotional attachment and symbolic ownership over a location by creating a personalized place name. This is common in sporting venues and political gathering sites (Tahir
Square/Maryr Square in Egypt). For all these cases, the rate of adoption (or non-compliance) and potential usage patterns would require longer observation periods.

In terms of co-occurrence, this can relate to a single event that simultaneously takes place at many locations, hence leading to a single event name replacing multiple place names. For example, the FIFA World Cup (single event) consists of football matches occurring in different stadiums (multiple official place names) of the host country. Additionally, multiple venues can host a New Years Eve party (single event) which all refer to the same celebration, albeit in different time zones and places.

In terms of street name phonetic similarity, there is room for development in the algorithm and also experiment range. Ranking refinement could be augmented with a more complete vowel assessment of the entire name/word and also the potential of considering syllable count as a factor of similarity. For example, the name pair of Addison/Adson produces a positive phonetic and n-gram result but is clearly distinct in the central vowel sound and syllable count. Additionally, the method of evaluating street suffix similarity is possibly too simplistic and requires on-the-ground polling/surveys to better represent the thought process of people. In terms of scoring distribution, there is also merit to look into weighting of the algorithm components based on cognitive assessment.

Syllable count, which is ignored in the current algorithm, should be taken seriously due to its influence on pronunciation (Kirchhoff 1996) and can be improved in current system. Phonetic algorithms indirectly provide a very rough representation of syllable count, with each digit/code akin to a single syllable, but with obvious limitations in accuracy. For now, the n-grams method indirectly compensates for the lack of vowel and syllable awareness by providing a measure spelling difference. Future work can look at research that has attempted to make syllables the sole unit of analysis in pattern matching (Gong and Chan 2006). In their case, syllable locations first have to be estimated, and then the minimum cost of converting one character group to another is calculated.

From the analysis of the results, a better algorithm can be produced by incorporating the additional variables discussed above (syllable count and central vowels) and by further tweaking the weightage of each component based on the human-annotated results. From a wider perspective, there is merit in looking into other street name characteristics. For
example, the Influence of generic particles in street names (Old Albert Road, Albert Road) does have an effect on perceptual similarity that could be parameterized. There is also motivation to further investigate how street names of varying cultural origins affect similarity clustering. This may reveal patterns in the similarities that occur for each culture type, and verify the speculated benefits of having mixed-culture street names in a city. Inversely, the language, dialect or accent of users can also lead to the same street name to be articulated in various ways. For example, different regions in Australia utilize vowel emphasis uniquely even though the language is the same (English) (Bradley 1989). At a larger picture this algorithm can also be tested on other place names besides streets.
BIBLIOGRAPHY


Author/s: Chan, Chun Keith

Title: Ambiguity resolution of place names: detecting event names associated with place and phonetic conflicts in street names

Date: 2014

Persistent Link: http://hdl.handle.net/11343/41005

File Description: Ambiguity resolution of place names: exploring place-event substitution and phonetic similarity in street name networks