Who Tweets? Deriving Demographic Information from Twitter

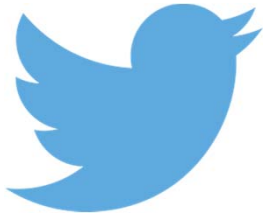
Dr Luke Sloan
Email: SloanLS@cardiff.ac.uk
Follow: @drlukesloan

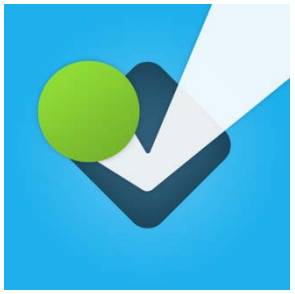
Collaborative Online Social Media ObServatory (COSMOS)

Luke Sloan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, Jeffrey Morgan, Rob Proctor & Omer Rana

Outline

- The Digital Revolution
- Emergence of Social Media
- Reflections on Representation
- ‘Naturally Occurring’ Data
- COSMOS
- Deriving Demographics
- Current Work





The Digital Revolution

- Every minute…
  - 48 hours of video uploaded to YouTube
  - 204,166,667 emails are sent
  - 2,000,000 search queries on Google
  - 217 users join the mobile web
  - 571 new website are created
  - 100,000+ tweets are made
  - 3,125 photos added to Flickr
  - 684,478 pieces of content shared on Facebook

Source: http://mashable.com/2012/06/22/data-created-every-minute/ [accessed April 2014]

It's all data!

Emergence of Social Media I

- Social Media is potentially a rich source of naturally occurring data on beliefs, attitudes, reactions and opinions

- For example, Twitter can be used for…
  - Brand tracking with sentiment (Scarfi 2012)
  - Predicting movie revenue (Asur & Huberman 2010)
  - Advance earthquake warning (Sakaki et al. 2010)
  - Predicting election results (Tumasjan et al. 2010)

Unlike traditional social science data collection, this can all be gathered for free…

Emergence of Social Media II

- The volume of data is phenomenal compared to a social survey…
  - ‘Spritzer’ at 1% approx. 3.5m tweets a day (free)
  - ‘Garden Hose’ at 10% approx 35m tweets a day (make a case)
  - ‘Fire Hose’ at 100% approx. 350m tweets a day (payment only)

Reflections on Representation

- But is Twitter representative?
  - Representative of who?
  - Population data for Twitter users (transactional)

- If Twitter data can be used to predict (*understand*) real world events then it should be representative at some level (but young people tweet, young people go to the cinema?)

- Much potential, but must be realised in a (social) scientific way

- At the heart of this is finding out who Tweets…

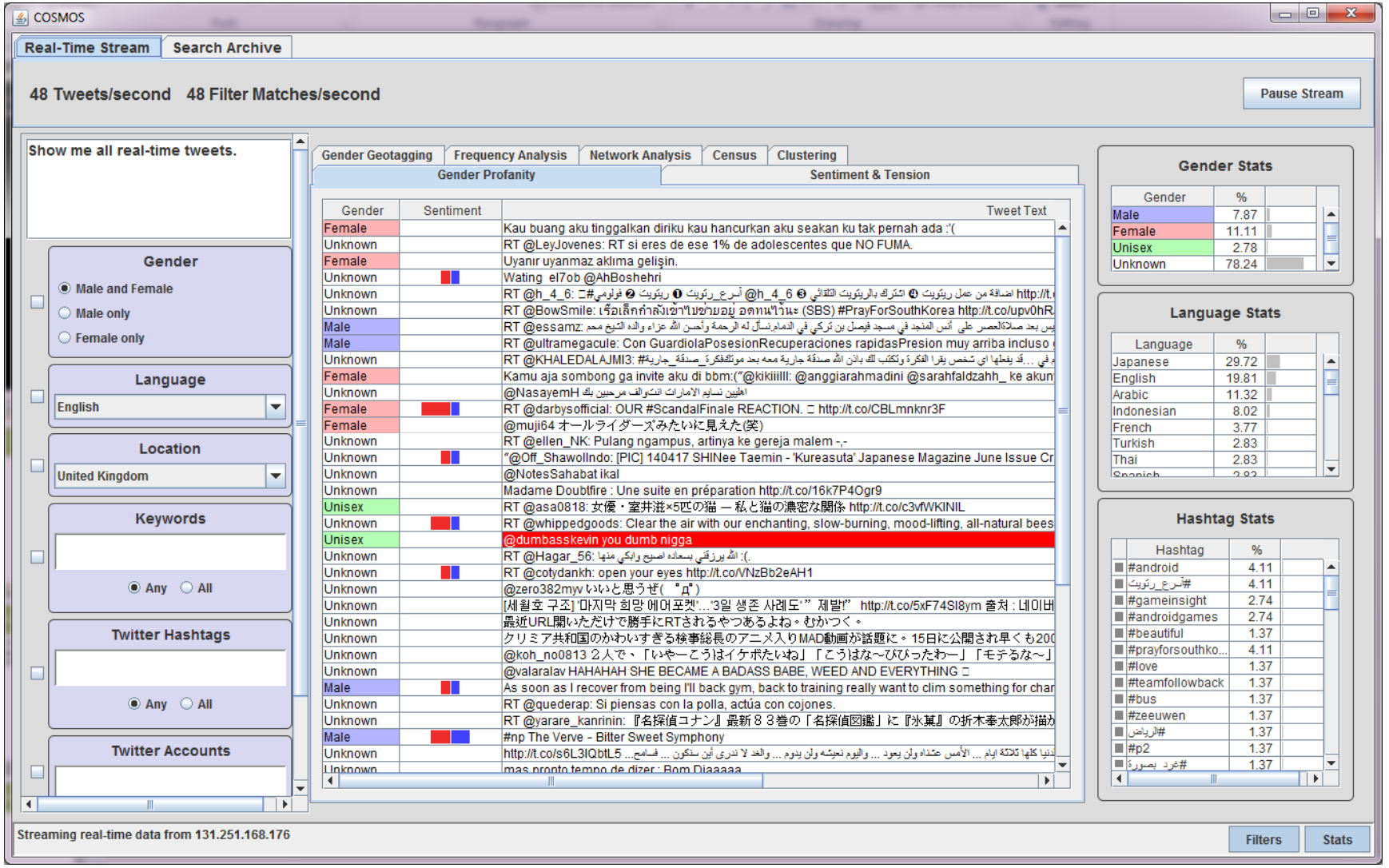
‘Naturally Occurring’ Data

- User generated content can be ‘data light’ (Mislove et al. 2011, Gayo-Avello 2012)
- Facebook is different because it stores baseline demographic information (Schwartz et al. 2013)
- Twitter has signatures, but nothing systematic
- When the data is not available we develop proxies, so why not for Twitter?

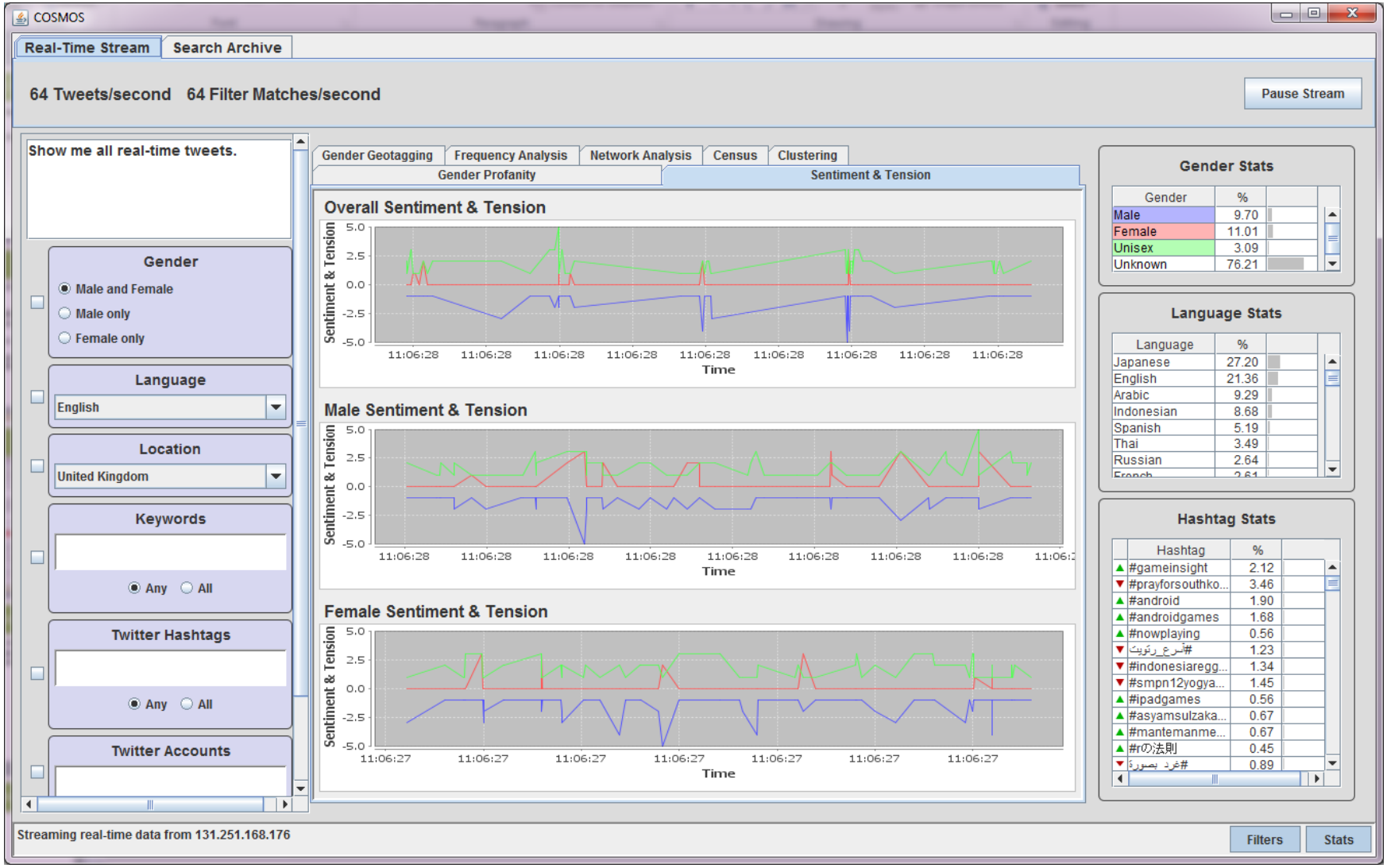
COSMOS I

- The Collaborative Online Social Media ObServatory (COSMOS) funded by…
- Develop a platform for data interoperability between:
  - naturally occurring data (e.g. social media)
  - curated data (e.g. social surveys)
  - administrative data (e.g. A&E admissions)
- Social media augments traditional social scientific investigation – it is not a surrogate! (Edwards et al. 2013)
- A key programme of work within COSMOS is on the derivation of demographic data from ‘signatures’ (aka *exhaust fumes*)

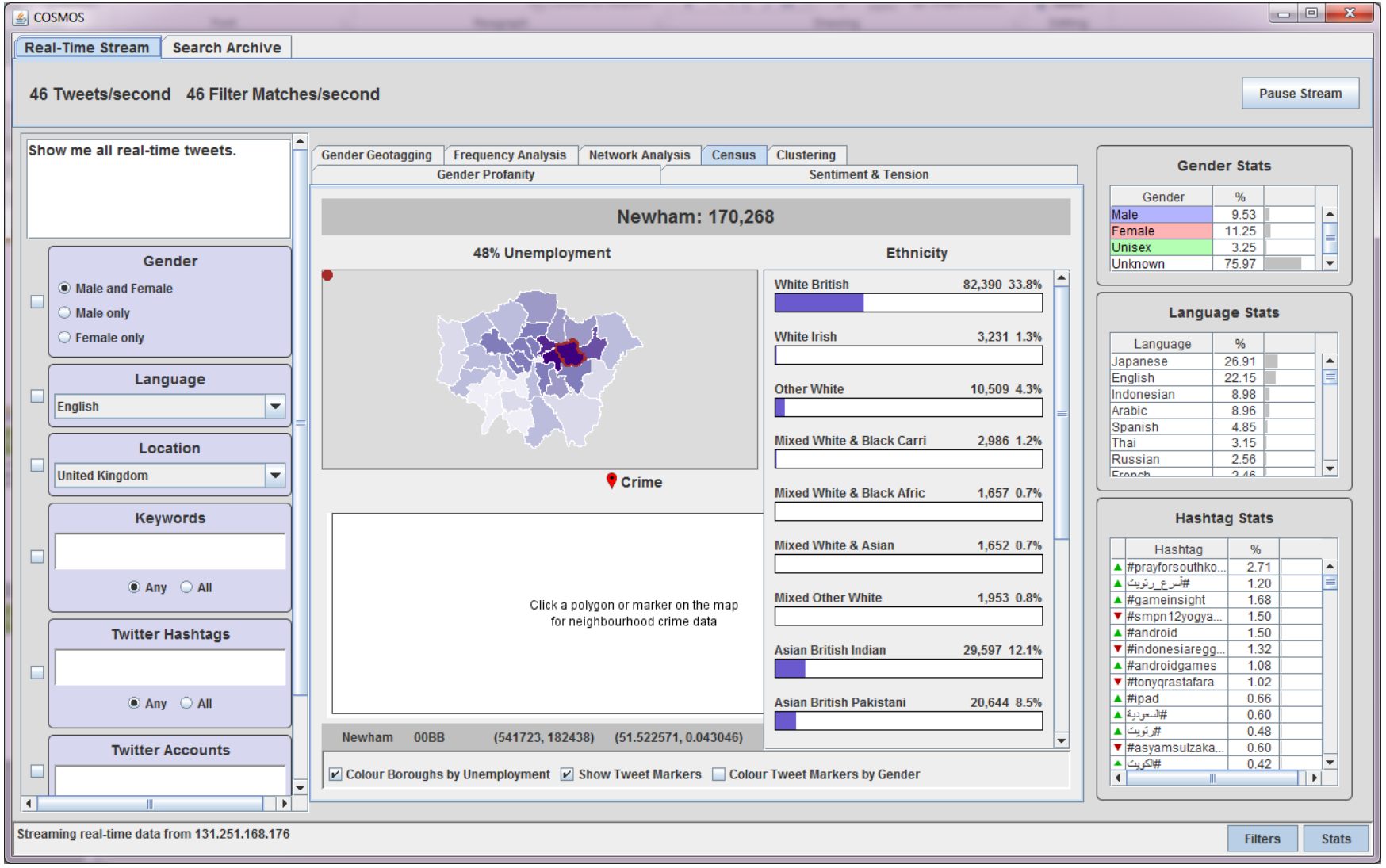
COSMOS II


COSMOS
Real-Time Stream
Search Archive
48 Tweets/second 48 Filter Matches/second
Pause Stream
Show me all real-time tweets.
Gender
Male and Female
Male only
Female only
Language
English
Location
United Kingdom
Keywords
Any All
Twitter Hashtags
Any All
Twitter Accounts
Gender Geotagging
Frequency Analysis
Network Analysis
Census
Clustering
Gender Profanity
Sentiment & Tension
Gender
Sentiment
Tweet Text
Female Kau buang aku tinggalkan diriku kau hancurkan aku seakan ku tak pernah ada :'(
Unknown RT @LeyJovenes: RT si eres de ese 1% de adolescentes que NO FUMA.
Female Uyanır uyanmaz aklıma gelişin.
Unknown Wating el7ob @AhBoshehri
Unknown RT @h_4_6: ⊐#فولومي ❷ ريتويت ❶ آسرع_ريتويت @h_4_6 ❸ اشترك بالريتويت التلقائي ❹ اضافة من عمل ريتويت http://t.
Unknown RT @BowSmile: เรือเล็กกำลังเข้าไปช่วยอยู่ อดทนไว้นะ (SBS) #PrayForSouthKorea http://t.co/upv0hR
Male RT @essamz: ليس بعد صلاةالعصر على أنس المنجد في مسجد فيصل بن تركي في الدمام نسأل له الرحمة وأحسن الله عزاء والده الشيخ محم
Male RT @ultramegacule: Con GuardiolaPosesionRecuperaciones rapidasPresion muy arriba incluso
Unknown RT @KHALEDALAJMI3: #جارية_صدقة_موتلفكرة_وتكتب لك باذن الله صدقة جارية معه بعد ... قد يفعلها اي شخص يقرا الفكرة وتكتب لك
Female Kamu aja sombong ga invite aku di bbm:("@kikiiiIII: @anggiarahmadini @sarahfaldzahh_ ke akun
Unknown @NasayemH اهلين نسايم الامارات التحوالف مرحبين بك
Female RT @darbysofficial: OUR #ScandalFinale REACTION. ⊐ http://t.co/CBLmnknr3F
Female @muji64 オールライダーズみたいに見えた(笑)
Unknown RT @ellen_NK: Pulang ngampus, artinya ke gereja malem -,-
Unknown "@Off_ShawolIndo: [PIC] 140417 SHINee Taemin - 'Kureasuta' Japanese Magazine June Issue Cr
Unknown @NotesSahabat ikal
Unknown Madame Doubtfire : Une suite en préparation http://t.co/16k7P4Ogr9
Unisex RT @asa0818: 女優・室井滋×5匹の猫 — 私と猫の濃密な関係 http://t.co/c3vfWKINIL
Unknown RT @whippedgoods: Clear the air with our enchanting, slow-burning, mood-lifting, all-natural bees
Unisex @dumbasskevin you dumb nigga
Unknown RT @Hagar_56: الله يرزقني بسعاده اصيح وابكي منها :(.
Unknown RT @cotydankh: open your eyes http://t.co/VNzBb2eAH1
Unknown @zero382myv いいと思うぜ(゜д゜)
Unknown [세월호 구조] '마지막 희망 에어포켓'…'3일 생존 사례도'" 제발!" http://t.co/5xF74Sl8ym 출처 : 네이버
Unknown 最近URL開いただけで勝手にRTされるやつあるよね。むかつく。
Unknown クリミア共和国のかわいすぎる検事総長のアニメ入りMAD動画が話題に。15日に公開され早くも200
Unknown @koh_no0813 2人で、「いやーこうはイケボたいね」「こうはな~びびったわー」「モテるな~」
Unknown @valaralav HAHAHAH SHE BECAME A BADASS BABE, WEED AND EVERYTHING ⊐
Male As soon as I recover from being I'll back gym, back to training really want to clim something for char
Unknown RT @quederap: Si piensas con la polla, actúa con cojones.
Unknown RT @yarare_kanrinin: 『名探偵コナン』最新８３巻の「名探偵図鑑」に『氷菓』の折木奉太郎が描か
Male #np The Verve - Bitter Sweet Symphony
Unknown http://t.co/s6L3lQbtL5 ... فسامح ... والغد لا ندري أين سنكون ... واليوم نعيشه ولن يدوم ... الأمس عشناه ولن يعود ... دنيا كلها ثلاثة ايام
Unknown mas pronto tempo de dizer : Bom Diaaaaa
Gender Stats
Gender % Male 7.87 Female 11.11 Unisex 2.78 Unknown 78.24
Language Stats
Language % Japanese 29.72 English 19.81 Arabic 11.32 Indonesian 8.02 French 3.77 Turkish 2.83 Thai 2.83
Hashtag Stats
Hashtag % #android 4.11 #آسرع_ريتويت 4.11 #gameinsight 2.74 #androidgames 2.74 #beautiful 1.37 #prayforsouthko... 4.11 #love 1.37 #teamfollowback 1.37 #bus 1.37 #zeeuwen 1.37 #الرياض 1.37 #p2 1.37 #غرد بصورة 1.37
Streaming real-time data from 131.251.168.176
Filters
Stats

COSMOS III


COSMOS
Real-Time Stream
Search Archive
64 Tweets/second 64 Filter Matches/second
Pause Stream
Show me all real-time tweets.
Gender
Male and Female
Male only
Female only
Language
English
Location
United Kingdom
Keywords
Any All
Twitter Hashtags
Any All
Twitter Accounts
Gender Geotagging
Frequency Analysis
Network Analysis
Census
Clustering
Gender Profanity
Sentiment & Tension
Overall Sentiment & Tension
Male Sentiment & Tension
Female Sentiment & Tension
Sentiment & Tension
Time
Gender Stats
Gender | %
Male | 9.70
Female | 11.01
Unisex | 3.09
Unknown | 76.21
Language Stats
Language | %
Japanese | 27.20
English | 21.36
Arabic | 9.29
Indonesian | 8.68
Spanish | 5.19
Thai | 3.49
Russian | 2.64
Hashtag Stats
Hashtag | %
#gameinsight | 2.12
#prayforsouthko... | 3.46
#android | 1.90
#androidgames | 1.68
#nowplaying | 0.56
#1.23
#indonesiaregg... | 1.34
#smpn12yogya... | 1.45
#ipadgames | 0.56
#asyamsulzaka... | 0.67
#mantemanme... | 0.67
#rの法則 | 0.45
0.89
Streaming real-time data from 131.251.168.176
Filters
Stats

COSMOS IV


COSMOS
Real-Time Stream
Search Archive
46 Tweets/second 46 Filter Matches/second
Pause Stream
Show me all real-time tweets.
Gender
Male and Female
Male only
Female only
Language
English
Location
United Kingdom
Keywords
Any All
Twitter Hashtags
Any All
Twitter Accounts
Gender Geotagging
Frequency Analysis
Network Analysis
Census
Clustering
Gender Profanity
Sentiment & Tension
Newham: 170,268
48% Unemployment
Ethnicity
Crime
Click a polygon or marker on the map
for neighbourhood crime data
White British 82,390 33.8%
White Irish 3,231 1.3%
Other White 10,509 4.3%
Mixed White & Black Carri 2,986 1.2%
Mixed White & Black Afric 1,657 0.7%
Mixed White & Asian 1,652 0.7%
Mixed Other White 1,953 0.8%
Asian British Indian 29,597 12.1%
Asian British Pakistani 20,644 8.5%
Newham 00BB (541723, 182438) (51.522571, 0.043046)
Colour Boroughs by Unemployment
Show Tweet Markers
Colour Tweet Markers by Gender
Gender Stats
Gender %
Male 9.53
Female 11.25
Unisex 3.25
Unknown 75.97
Language Stats
Language %
Japanese 26.91
English 22.15
Indonesian 8.98
Arabic 8.96
Spanish 4.85
Thai 3.15
Russian 2.56
Hashtag Stats
Hashtag %
#prayforsouthko... 2.71
#1.20
#gameinsight 1.68
#smpn12yogya... 1.50
#android 1.50
#indonesiaregg... 1.32
#androidgames 1.08
#tonyqrastafara 1.02
#ipad 0.66
0.60
0.48
#asyamsulzaka... 0.60
0.42
Streaming real-time data from 131.251.168.176
Filters
Stats

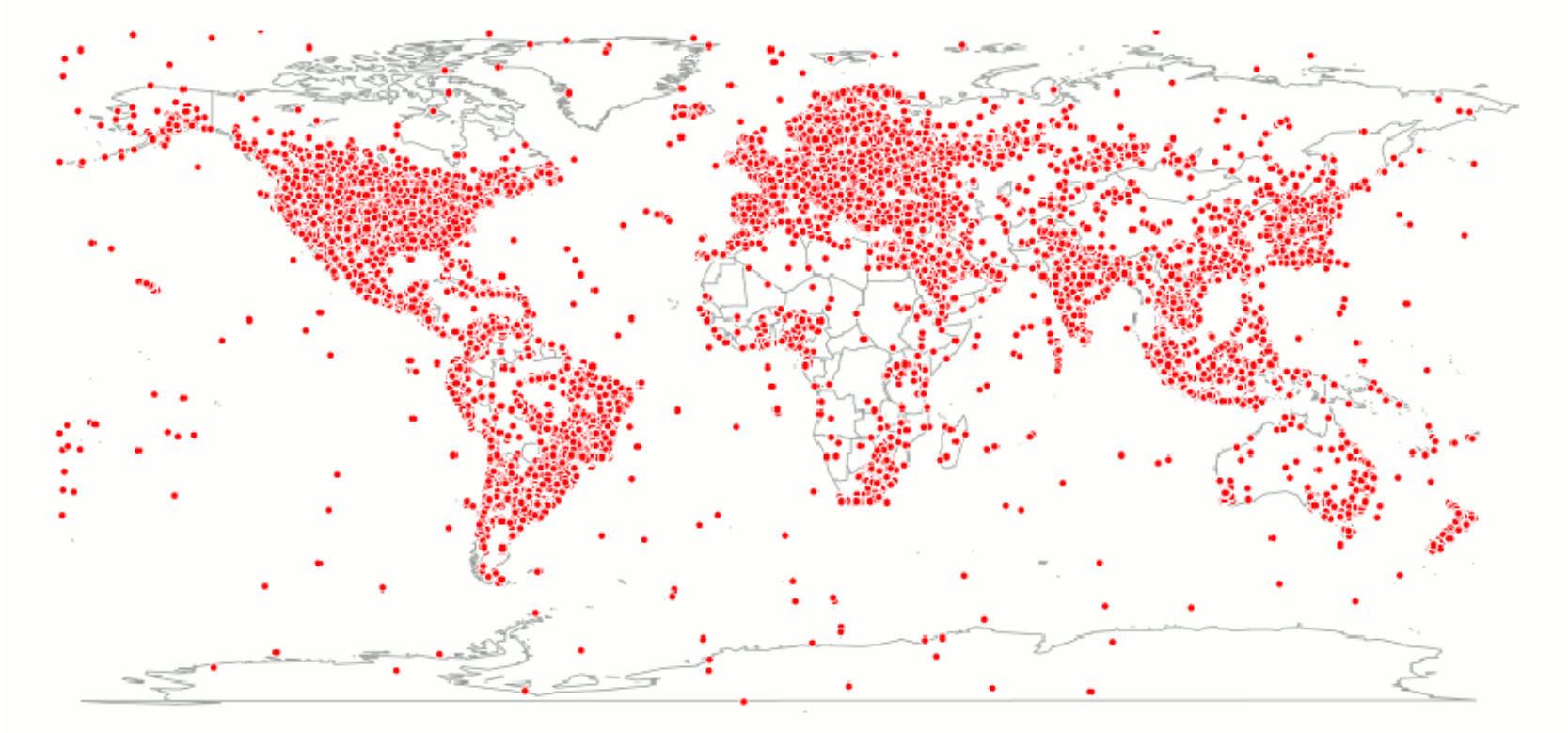
Deriving Demographics

- Development of tools embedded within the COSMOS platform to identify signatures of demographic characteristics (Sloan et al. 2013)
- Location
- Gender
- Language

Deriving Demographics: Location I

- Three primary sources of location:
  - User profile information
  - Content of tweets (inc. ‘mundane geography’)
  - Geo-tagged tweets
- Geo-tagged tweets are the gold standard
- Allows us to locate people at the time they tweeted in existing geographies (output area level!)
- RQ: do people tweet about crime in high crime areas?

Deriving Demographics: Location II



Source: Sloan et al. 2013

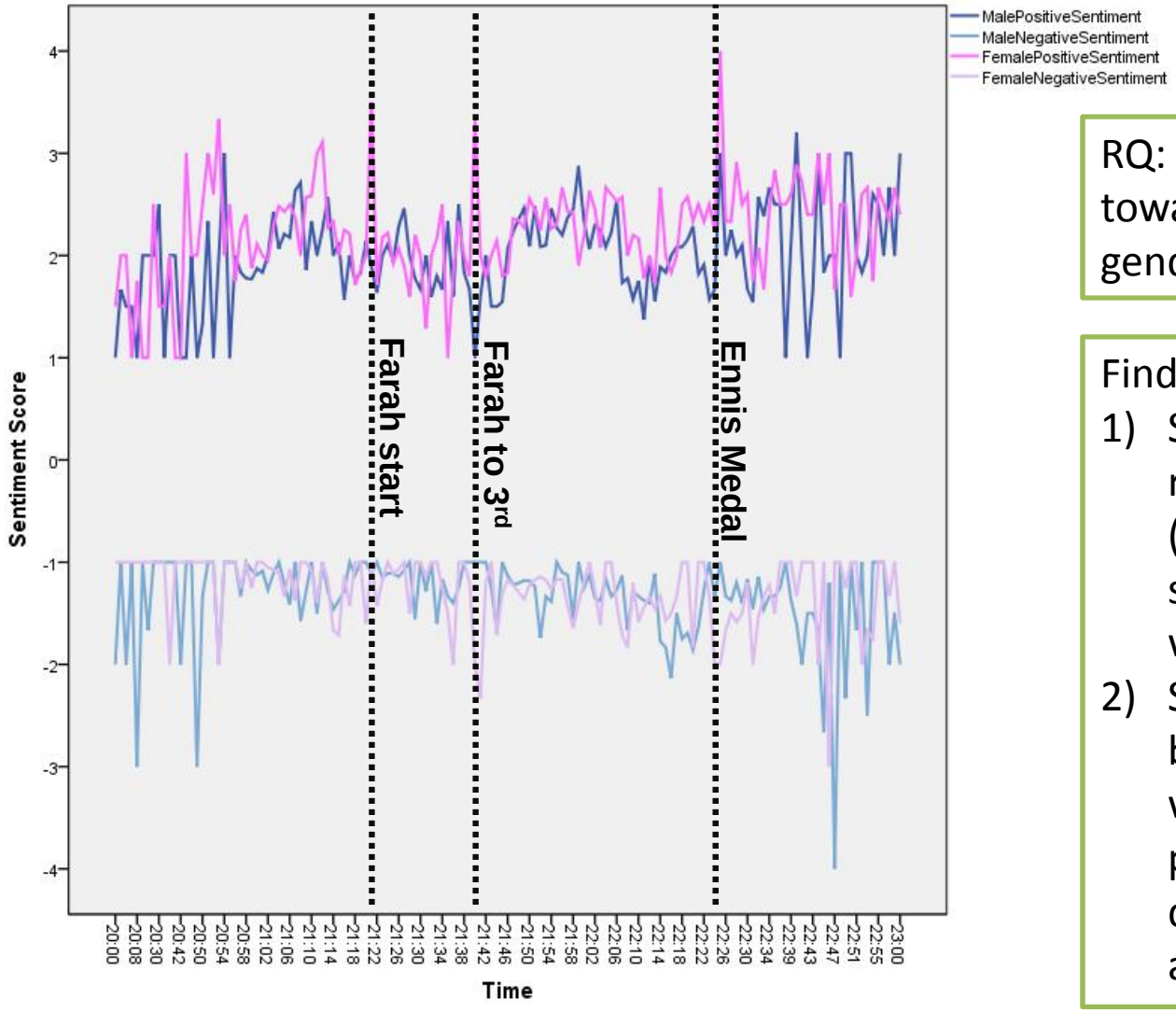
Deriving Demographics: Language

- Two methods of identifying language:
  - The language of the Twitter interface
  - The language of the Tweet (Java library)
- Detecting language is important for efficiency of other analytical tools (e.g. sentiment analysis)
- 40% of content is in English
- RQ: do spatial patterns of language use recorded on the Census correspond with Twitter maps?
- Note that ‘hard to reach groups’ may use Twitter…

Deriving Demographics: Gender I

- Use the name field of the Twitter profile
- Clean the data to extract a first name and compare against a large database of first names
- Important to categorise ‘unisex’ and ‘unknown’
- Of those we could identify: 48.8% male and 51.2% female… exactly the same as the 2011 Census

Deriving Demographics: Gender II



RQ: How does sentiment towards Team GB differ by gender?

Findings:

1) Sentiment peaks reflect real world events (relationship between social media and real world)
2) Sentiment differs between men and women (difference is so pronounced that gender detection method appears to work)

Current Work

- Identifying age from signature data:
  - Preliminary analysis suggests usable age data for 0.35% of Twitter users
  - Note that 0.35% of 645m is 2.25m (approx 40% of which is English language)

- Identify occupation from signature data:
  - Linked to SOC2010 codes
  - Enables allocation into NS-SEC groups

Discussion

References

Asur & Huberman (2010) Predicting the Future with Social Media, *Social Computing Lab,* HP Labs in Palo Alto

Edwards et al. (2013) Computational social science and methodological innovation: surrogacy, augmentation or reorientation?, *International Journal of Social Research Methods*, 16:3

Gayo-Avello (2012) I wanted to Predict Elections with Twitter and all I got was this Lousy Paper: A Balanced Survey on Election Prediction using Twitter Data, *Department of Computer Science*, University of Oviedo Spain

Mislove et al. (2011) Understanding the demographics of Twitter users, *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*

Sakaki et al. (2010) Earthquake shakes Twitter users: real-time event detection by social sensors, *presented at WWW 2010*, Raleigh, NC USA

Scarfi (2012) Social Media and the Big Data Explosion, *Forbes* (www.forbes.com/sites/onmarketing/2012/06/28/social-media-and-the-big-data-explosion/)

Schwartz et al. (2013) Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach, *PLOS ONE*, 8:9 (DOI: 10.1371/journal.pone.0073791)

Sloan et al. (2013) Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter, *Sociological Research Online*, 18:3 (http://www.socresonline.org.uk/18/3/7.html)

Tumasjan et al. (2010) Predicting elections with Twitter: what 140 characters reveal about political sentiment, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*

Who Tweets?

Dr Luke Sloan
Email: SloanLS@cardiff.ac.uk
Follow: @drlukesloan

Collaborative Online Social Media ObServatory (COSMOS)

Luke Sloan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, Jeffrey Morgan, Rob Proctor & Omer Rana