My experiences with Social Media Analysis so far

Fabio Giglietto
(fabio.giglietto@uniurb.it)

1506
UNIVERSITÀ
DEGLI STUDI
DI URBINO
CARLO BO

DISCUM
DIPARTIMENTO DI
SCIENZE DELLA COMUNICAZIONE
E DISCIPLINE UMANISTICHE

Dealing with platforms APIs

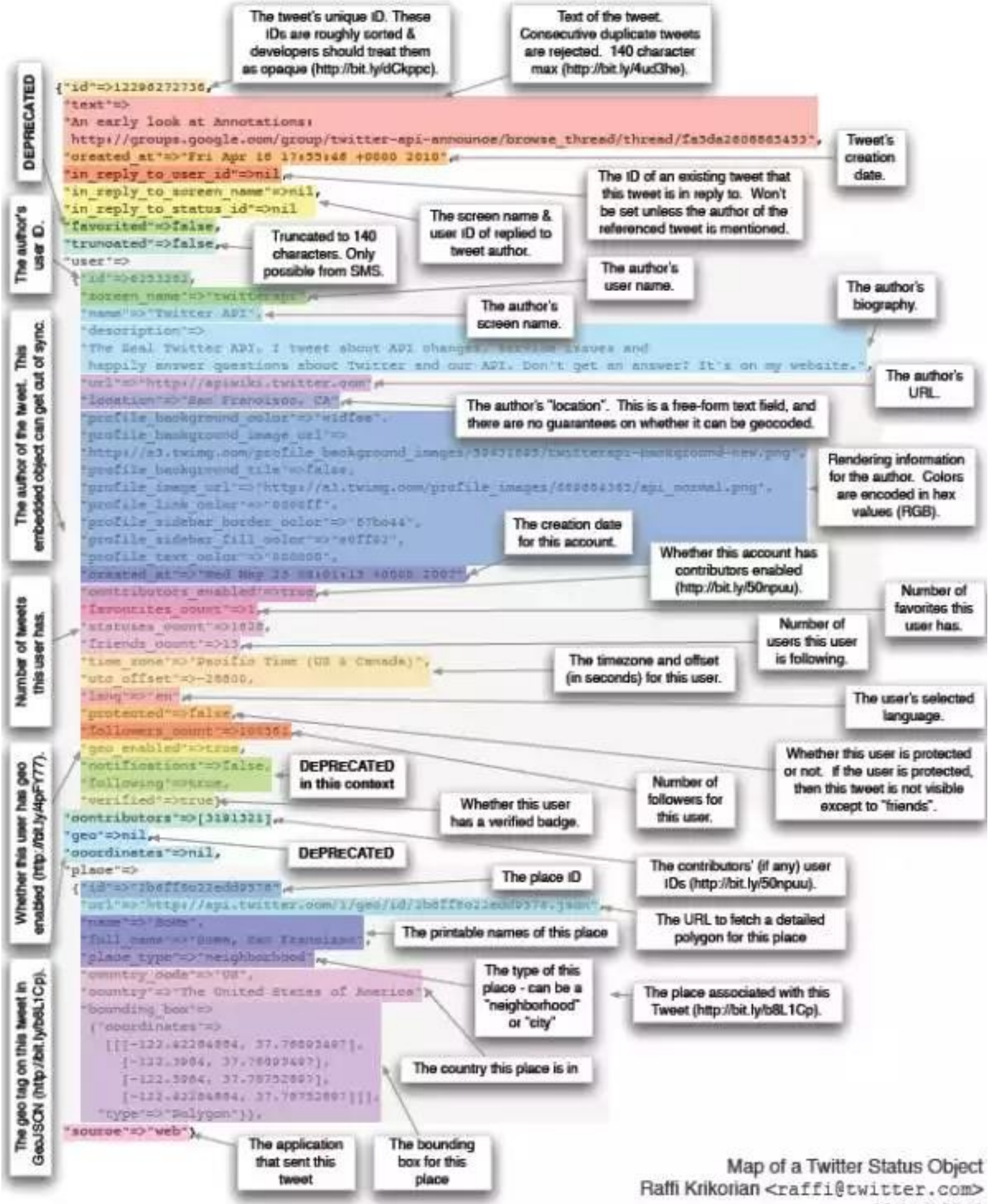
- Facebook
  - Graph API
  - Apps
  - Public Feed API & Keyword Insights API
- Twitter
  - Search API
  - Streaming API
    - DMI-TCAT, StreamR
  - Firehose
    - GNIP (Sifter), DataSift
    - DiscoverText, TweetReach

The dataset

- From August 30th, 2012 to June 30th, 2013;
- Over 3 million tweets created by 270,000 unique contributors;
- containing the official #hashtags of
  - 11 political talk shows;
  - the 6th Italian edition of “X Factor”.
- From GNIP/Twitter *firehose* (no search or Streaming API);

Main issues encountered

- Twitter Free APIs provide “not good enough samples”, but purchasing tweets is expensive;
- Dealing with and managing a large dataset in JSON format;
- Data Analysis with R;
- Moving from big to “deep data”: limits of sampling and possible alternatives.



```
{"id"=>12296272736,
 "text"=>
 "An early look at Annotations:
  http://groups.google.com/group/twitter-api-announce/browse_thread/thread/fa3da2608865453",
 "created_at"=>"Fri Apr 16 17:55:46 +0000 2010",
 "in_reply_to_user_id"=>nil,
 "in_reply_to_screen_name"=>nil,
 "in_reply_to_status_id"=>nil
 "favorited"=>false,
 "truncated"=>false,
 "user"=>
  {"id"=>6253282,
   "screen_name"=>"twitterapi",
   "name"=>"Twitter API",
   "description"=>
   "The Real Twitter API. I tweet about API changes, service issues and
    happily answer questions about Twitter and our API. Don't get an answer? It's on my website.",
   "url"=>"http://apiwiki.twitter.com",
   "location"=>"San Francisco, CA",
   "profile_background_color"=>"c1dfee",
   "profile_background_image_url"=>
   "http://a3.twimg.com/profile_background_images/59931895/twitterapi-background-new.png",
   "profile_background_tile"=>false,
   "profile_image_url"=>"http://a3.twimg.com/profile_images/689684365/api_normal.png",
   "profile_link_color"=>"0000ff",
   "profile_sidebar_border_color"=>"87bc44",
   "profile_sidebar_fill_color"=>"e0ff92",
   "profile_text_color"=>"000000",
   "created_at"=>"Wed May 23 06:01:13 +0000 2007",
   "contributors_enabled"=>true,
   "favorites_count"=>1,
   "statuses_count"=>1628,
   "friends_count"=>13,
   "time_zone"=>"Pacific Time (US & Canada)",
   "utc_offset"=>-28800,
   "lang"=>"en",
   "protected"=>false,
   "followers_count"=>100581,
   "geo_enabled"=>true,
   "notifications"=>false,
   "following"=>true,
   "verified"=>true},
 "contributors"=>[3191321],
 "geo"=>nil,
 "coordinates"=>nil,
 "place"=>
  {"id"=>"2b6ff8c22edd9576",
   "url"=>"http://api.twitter.com/1/geo/id/2b6ff8c22edd9576.json",
   "name"=>"SoMa",
   "full_name"=>"SoMa, San Francisco",
   "place_type"=>"neighborhood",
   "country_code"=>"US",
   "country"=>"The United States of America",
   "bounding_box"=>
    {"coordinates"=>
      [[[-122.42284884, 37.76893497],
        [-122.3964, 37.76893497],
        [-122.3964, 37.78752897],
        [-122.42284884, 37.78752897]]],
     "type"=>"Polygon"}},
 "source"=>"web"}
```

The tweet's unique ID. These IDs are roughly sorted & developers should treat them as opaque (http://bit.ly/dCkppc).

Text of the tweet. Consecutive duplicate tweets are rejected. 140 character max (http://bit.ly/4ud3he).

Tweet's creation date.

DEPRECATED

The ID of an existing tweet that this tweet is in reply to. Won't be set unless the author of the referenced tweet is mentioned.

The screen name & user ID of replied to tweet author.

Truncated to 140 characters. Only possible from SMS.

The author's user ID.

The author's user name.

The author's screen name.

The author's biography.

The author of the tweet. This embedded object can get out of sync.

The author's URL.

The author's "location". This is a free-form text field, and there are no guarantees on whether it can be geocoded.

Rendering information for the author. Colors are encoded in hex values (RGB).

The creation date for this account.

Whether this account has contributors enabled (http://bit.ly/50npuu).

Number of favorites this user has.

Number of tweets this user has.

Number of users this user is following.

The timezone and offset (in seconds) for this user.

The user's selected language.

Whether this user is protected or not. If the user is protected, then this tweet is not visible except to "friends".

Number of followers for this user.

Whether this user has geo enabled (http://bit.ly/4pFY77).

DEPRECATED in this context

Whether this user has a verified badge.

The contributors' (if any) user IDs (http://bit.ly/50npuu).

DEPRECATED

The place ID

The URL to fetch a detailed polygon for this place

The printable names of this place

The type of this place - can be a "neighborhood" or "city"

The place associated with this Tweet (http://bit.ly/b8L1Cp).

The geo tag on this tweet in GeoJSON (http://bit.ly/b8L1Cp).

The country this place is in

The application that sent this tweet

The bounding box for this place

Map of a Twitter Status Object
Raffi Krikorian <raffi@twitter.com>
18 April 2010

Predicting TV Audience

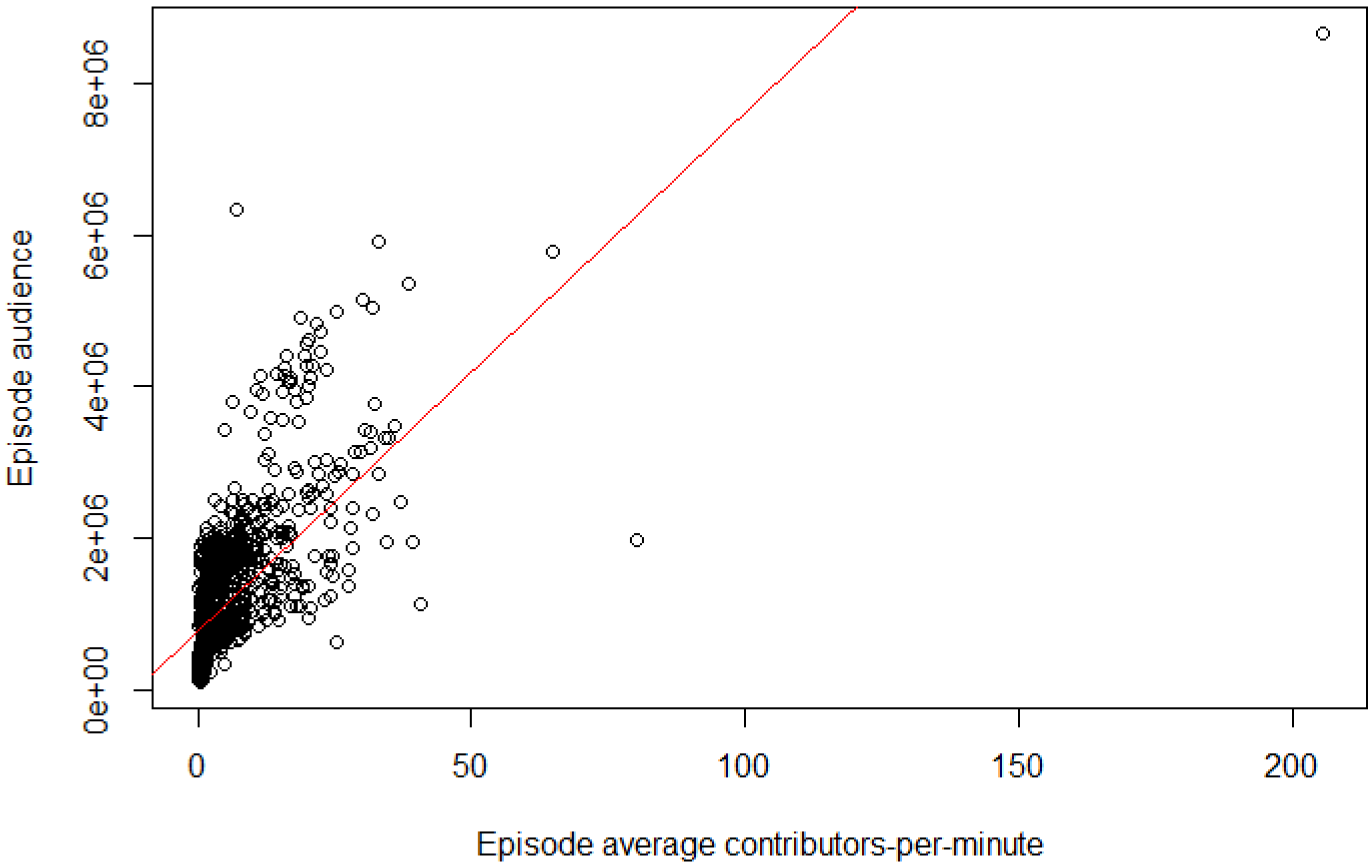
Dataset preparation

1. Subset of Tweets (1) created during the on air time of the episodes (+15 mins) and (2) containing the corresponding program #hashtag (n= 1,881,873);
2. 1,077 aired episodes with respective average audience and rating as estimated by Auditel;
3. Twitter metrics for each episode (Tweets, contributors, reach, ReTweet, Reply, Tweet-per-minute, contributors-per-minute).

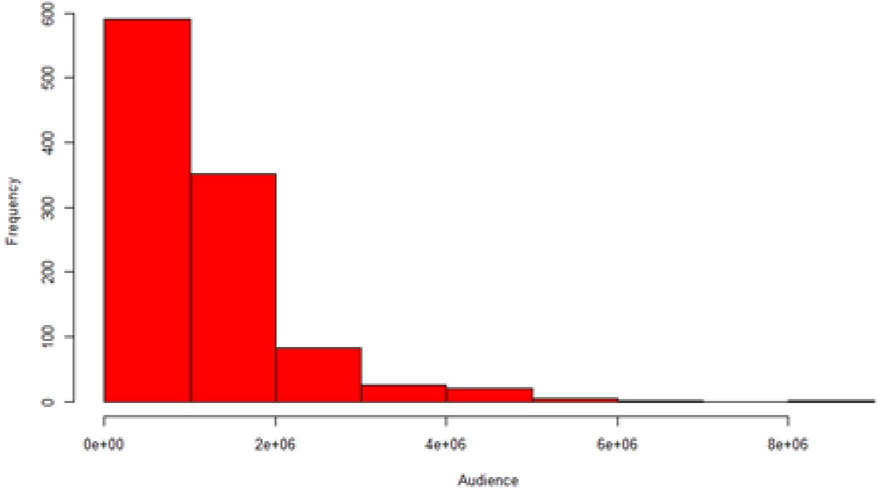
Correlation coefficients

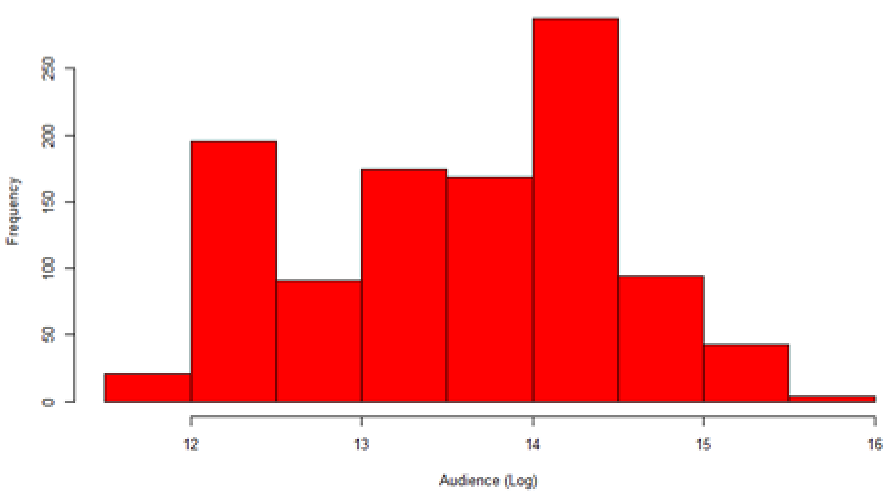
	Audience	**n**	**p**
Tweet	.54	1077	< .01
Contributors	.64	1077	< .01
Reach	.51	1077	< .01
ReTweet	.54	1077	< .01
Reply	.6	1077	< .01
Tweet-per-minute (TPM)	.57	1077	< .01
Contributors-per-minute (CPM)	.67	1077	< .01

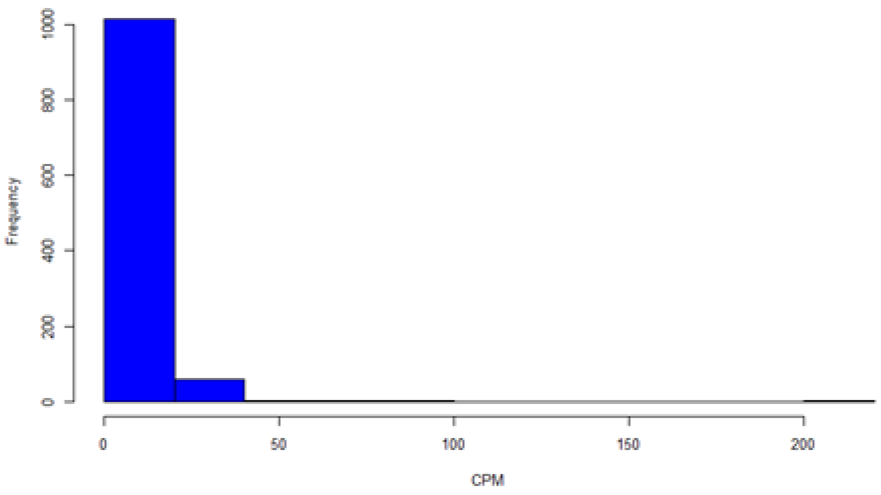
Audience ~ CPM

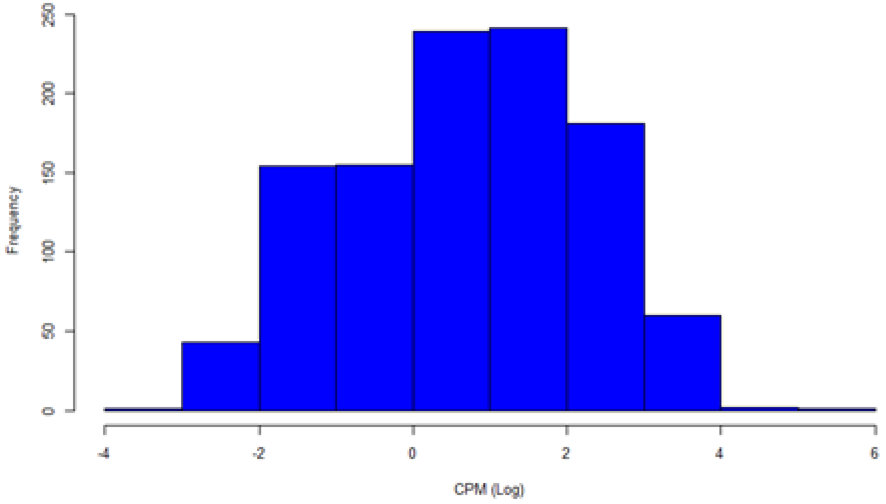


Loglinear transformation

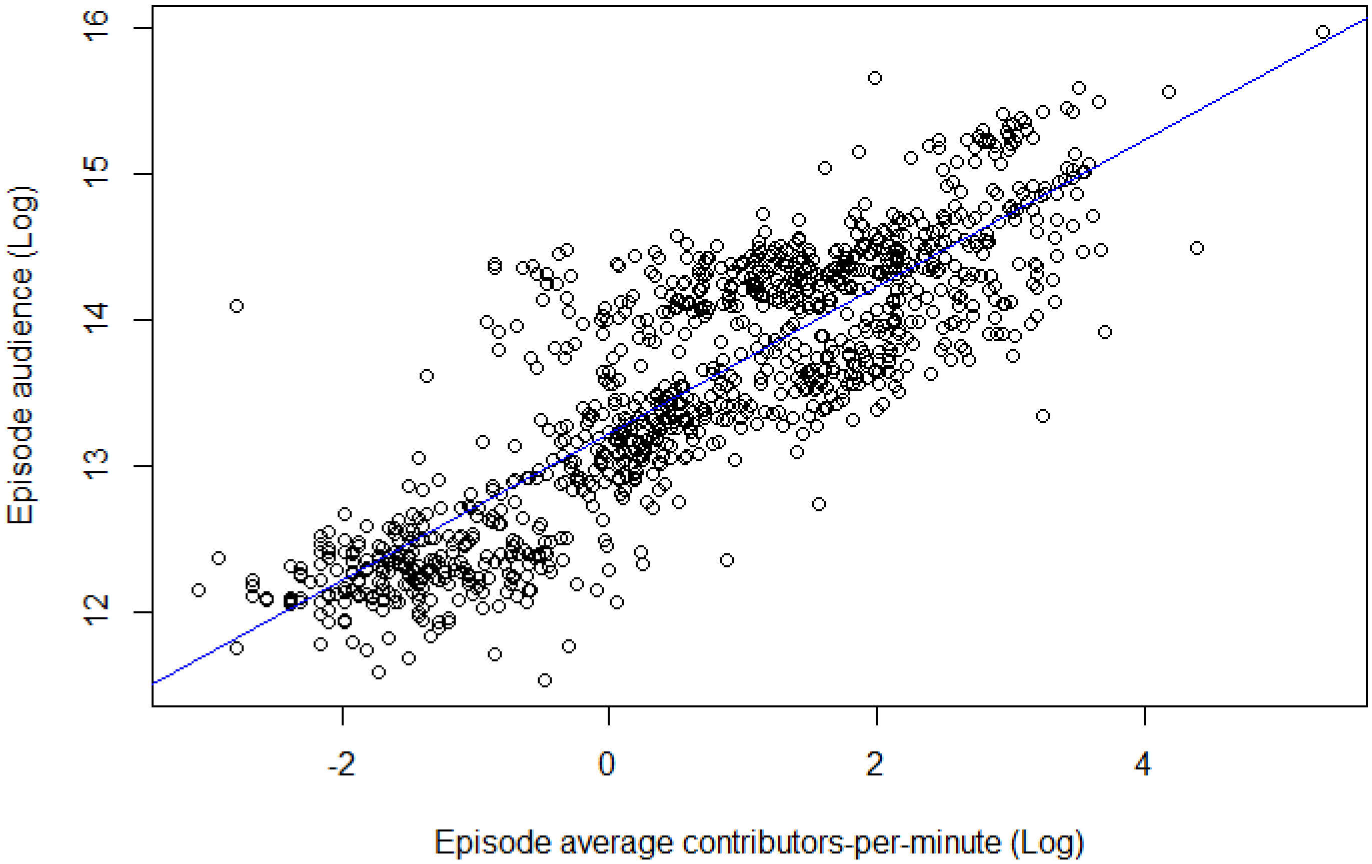








Log(Audience) ~ Log(CPM)



Correlations

	Audience	n	p
Tweet	.54	1077	< .01
Contributors	.64	1077	< .01
Reach	.51	1077	< .01
ReTweet	.54	1077	< .01
Reply	.6	1077	< .01
Tweet-per-minute (TPM)	.57	1077	< .01
Contributors-per-minute (CPM)	.67	1077	< .01
Log (CPM)	.86	1077	< .01

Results (1/3)

1. Over the eight different metrics tested, the observed correlation coefficient with the audience was > 0.5;
2. The rate of Tweet per minute (TPM) and contributors per minute (CPM) correlate remarkably well with audience (when log transformed respectively r=0.83 and 0.86) thus suggesting a strong non linear correlation;

Results (2/3)

- A multiple regression model based on the (1) average audience of previously aired episodes, (2) CPM and (3) networked publics variable*, explained 96% of the variance in the audience;
- Taking all other variables constant, we expect an increase of 0.37% in audience for an increase of 1% in average CPM;

* representing the inclination of the audience base of a show to contribute to the conversation with the official hashtag while the show is on air

Results (3/3)

- A linear model based on TPM only seems to be unable to efficiently predict the episode audience;
- Metrics extrapolated from Twitter activity could be successfully used to increase the precision of the prediction based on average past audience.

Understanding TV Genre Engagement and Willingness to Speak Up

SERVIZIO PUBBLICO

X FACTOR

Research Questions

- **RQ1.** What are specific moments of political talk show ”Servizio Pubblico” as well as of the entertainment Tv format “XFactor” that trigger audience engagement?

- **RQ2.** What are the most significant elements of continuity or discontinuity between these Tv show-based active audience regarding contents or communicative styles?

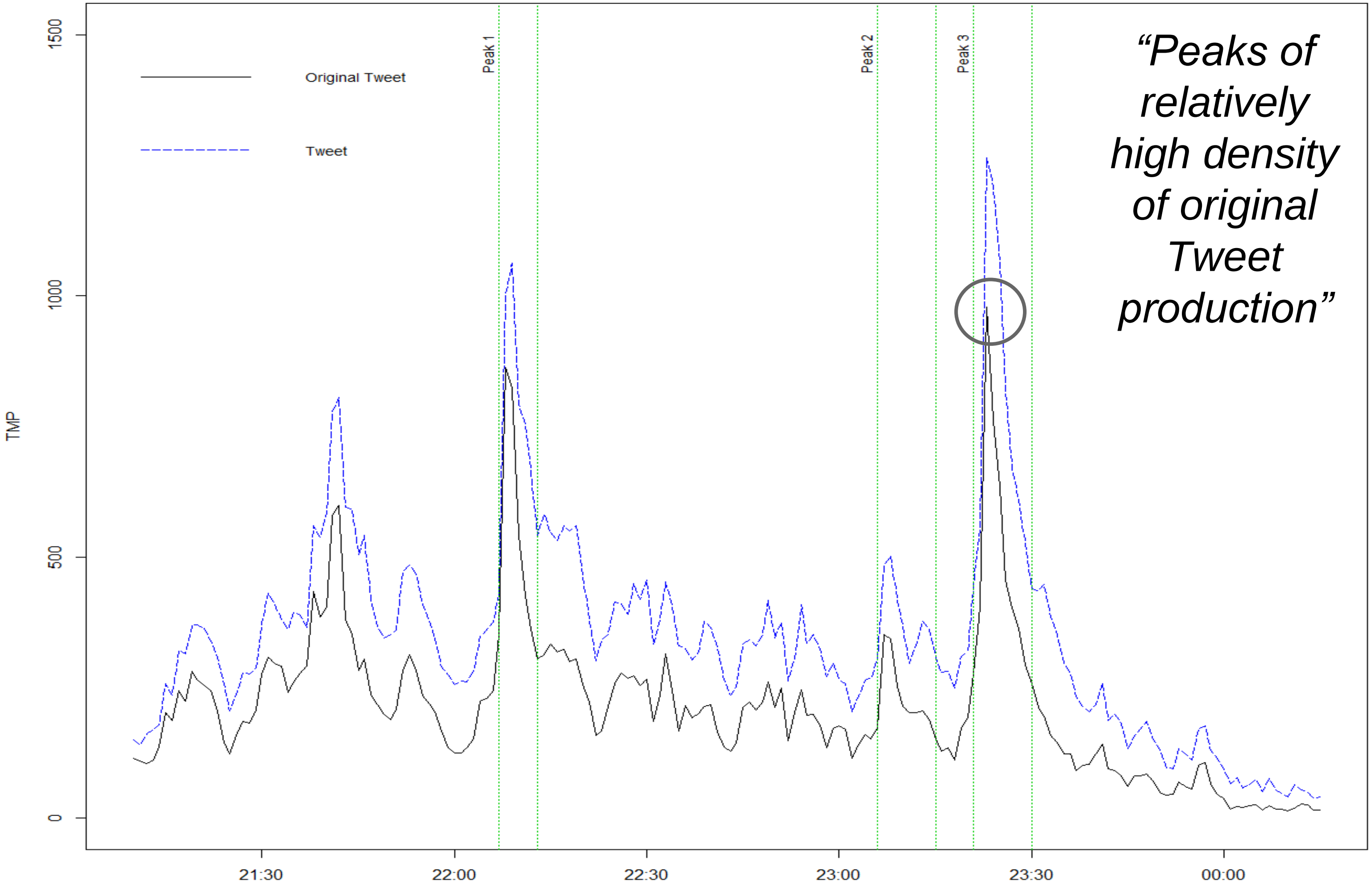
Dataset

2012/2013 Tv season	Official Hashtags	Episodes	Tweet	Unique Contributors
X Factor 6	#xf6	9	772,018	83,989
Servizio Pubblico	#serviziopubblico	28	611,396	96,911

	Minutes	Tweet	RT (%)	Replies (%)	Original Tweets (%)	Tweet Per Minute (tweet)
X Factor 6	221,780	772,018	31	6	62	3.48
Servizio Pubblico	439,201	611,396	41	4	55	1.39

	Episodes	Avg. Tweet/episode (SD)	Avg. TPM/episode (SD)
X Factor 6	9	62,489.33 (9,820.23)	337.78 (53.08)
Servizio Pubblico	28	16,934.54 (26,698.25)	99.61 (158.76)

Peaks of Twitter Engagement (PTE)



“Peaks of relatively high density of original Tweet production”

Peak Analysis: Procedure & Codeset

TV scene summary	Routine of the show	Luhmann's media system "selector" criteria	Tweet	RT	@replies	Original tweet	TPM

RQ1 Data Analysis (1/3)

	Peaks (N)	**Surprise** - *break with existing expectations* (%)	**Suspense** - *space of limited possibilities kept open* (%)
X Factor 6	16	**50**	**56.2**
Servizio Pubblico	39	**48.7**	5.1

RQ1 Data Analysis (2/3)

	Peaks (N)	Avg. TPM	Avg. Original Tweets (%)	Avg. RT (%)	Avg. Replies (%)
X Factor 6	16	**590.2**	**70**	25	5
Servizio Pubblico	39	248.31	63	**33**	4

RQ1 Data Analysis (3/3)

Servizio Pubblico

	Peaks				
Routine of the show	N	%	AVG TPM	% RT	% tweet originali
Talk show	31	79	**231.65**	33	63
Editorial by Marco Travaglio	5	13	**397.2**	39	59
Pre-recorded video	4	10	103.65	40	57
Member of the studio audience speaking	3	8	168.37	31	**64**
Poll results	2	5	118.69	39	56
Interview	1	2	68.43	41	56

X Factor 6

	Peaks				
Routine of the show	N	%	AVG TPM	% RT	% tweet originali
Contestant’s performance	4	25	**707.94**	20	**74**
Judge's comment	2	12	**695.38**	31	**75**
Results I part	3	18	602.76	31	70
Results II part	1	6	325.75	24	71
“Tilt”	2	12	403.98	25	69
Favorite song performance	1	6	352.75	31	71
A cappella performance	1	6	416	34	61
Elimination	6	37	**612.19**	26	70

Research Questions

- **RQ1.** What are specific moments of political talk show ”Servizio Pubblico” as well as of the entertainment Tv format “XFactor” that trigger audiences engagement?
- **RQ2.** What are the most significant elements of continuity or discontinuity between these Tv show-based active audiences regarding contents or communicative styles?
  - **RQ2a.** Do people tend to delegate and/or cover up the expression of opinions, when the show deals with politics rather than entertainment?
  - **RQ2b.** Is there a significant difference in the amount of Twitter expressions combined with informations when looking at peaks with high or low percentages of original tweets?

Peaks sampling

#serviziopubblico				
Peak id	Tweet	Original tweets	Original tweets:tweets (%)	Low OT %
9	466	232	50	TRUE
7	1,253	642	51	TRUE
29	519	380	73	FALSE
25	1,090	833	76	FALSE

#XF6				
Peak id	Tweet	Original tweets	Original tweets:tweets (%)	Low OT %
15	2,281	2,281	61	TRUE
16	4,823	4,823	63	TRUE
1	2,854	2,161	76	FALSE
10	1,665	1,279	77	FALSE

Content Analysis Codebook

	#XF6	#ServizioPubblico
Information	the one knocked out tonight was Nice #XF6	"We want to work but also to live" #ilva #serviziopubblico
Opinion	#XF6 Ics smashes guys!!!	good speeches until now at #serviziopubblico
Opinion (as joke)	Ics blends with the stage floor #sapevatelo #XF6	#serviziopubblico #cacciari is ready for fighting, it’s great!!!
Attention seeking	#XF6 ok, i’m going to turn off the PC and enjoy the voice of #Chiara...	I wonder what #serviziopubblico became?
Emotion	#Chiara AAAAAAAAAAAAAAAAAAAAA #XF6 ❤💛❤💛❤💛❤💛❤💛❤💛❤💛❤💛❤💛	Fuck off Cacciari!!! #serviziopubblico
Interaction	Please, take away the microphone from #Chiara #XF6 #xfactor6	#Madia go away. You learned the speech by heart!! #serviziopubblico

RQ2a Data Analysis

	% of all coded tweets (N=13,189)	% in #serviziopubblico (N=1,977)	% in #xf6 (N=11,212)
Information	21	27	15
Opinion	44	39	47
Opinion (as joke)	18	**25**	**11**
Emotion	3	**3**	**33**
Attention seeking	5	9	7
Interaction	11	12	15
Non coded	7	4	6
Total opinion	62	64	58
Information & opinion	7	**10**	**4**

Chi square were calculated for tweets belonging to #servizio pubblico and #xf6. The association between formats and all the categories is statistically significant (two-tailed P values < .001).

RQ2b Data Analysis

	#serviziopubblico	
	Tweets in peaks with LOW Original Tweets (N=909)	Tweets in peaks with HIGH Original Tweets (N=1,068)
Information + opinion (%)	**13***	**7***

	#XF6	
	Tweets in peaks with LOW Original Tweets (N=3,699)	Tweets in peaks with HIGH Original Tweets (N=7,513)
Information + opinion (%)	5	4

Chi square were calculated for tweets in low and high originali tweets. * p < .05, ** p < .01, *** p> .001

Conclusions (1/2)

1. *Framing effect* of Tv formats on Twitter active audiences
2. In both political and talent show, peaks of Twitter engagement are generated by surprise;
3. Suspense is a key engagement for talent show;
4. Original tweets are more frequent during talent show than political talk show thus suggesting a form of coaching participation. When an audience’s peer is on screen (member of in-studio audience or contestant) original tweets are also more frequent;

Conclusions (2/2)

5. Opinions are more frequently expressed as a joke or linked to information during political talk-shows rather than talent-shows;
6. In political talk-show, peaks with less original tweets also have more tweets coded as “information+opinion”;
7. Tweets expressing emotions are frequent during talent show and rare during political talk-shows.

Workshop on Analysing Twitter Social TV using R

Fabio Giglietto (fabio.giglietto@uniurb.it)

1506
UNIVERSITÀ DEGLI STUDI DI URBINO CARLO BO

DISCUM
DIPARTIMENTO DI SCIENZE DELLA COMUNICAZIONE E DISCIPLINE UMANISTICHE

Summary

1. Brief introduction to R and R Studio;
2. Getting the data from Twitter Streaming API;
3. Dataset Download;
4. Structure of a Twitter data-frame;
5. Counting unique contributors;
6. Counting RT and @replies;
7. Creating a timeline chart;
8. Detecting breakouts and peaks;
9. Setup for a content analysis of tweets in a peak.