EUROPEAN CENTRAL BANK

EUROSYSTEM

Workshop on using big data for forecasting and statistics

Monday, 7 and Tuesday, 8 April 2014
European Central Bank, Eurotower
Frankfurt am Main

EUROPEAN CENTRAL BANK

EUROSYSTEM

Paola Cerchiello
University of Pavia, Italy
Dep. Economics and Management
speaker 1

Paolo Giudici
University of Pavia, Italy
Dep. Economics and Management
speaker 2

How to Measure the Quality of Financial Tweets

SUMMARY

- Big data may be a useful source of information for financial forecasts.
- However, it cannot be used "as it is found"
- A measure of quality of its information content is needed.
- Here we focus on financial tweets and propose a statistical method to asses their quality, leading to a complete ranking of different twitter sources
- The method is quite effective and has being tested using library twitteR from the R sofwtare

EUROPEAN CENTRAL BANK

EUROSYSTEM

BACKGROUND

The *h* index:

"a scientist has index *h* if h of his or her N_p papers have at least *h* citations each and the other (N_p-*h*) papers have ≤ *h* citations each" (Hirsch, 2005).

Consider the ordered sample of retweets $\{X_{(t)}\}$, that is $X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}$.

The *h* index can be defined as follows:

$$h = \max\{t: X_{(t)} \geq t\}$$

Pro's: the h index is easy to calculate and to interpret

Con's: the h index as such discards the stochastic nature of the underlying data

EUROPEAN CENTRAL BANK

EUROSYSTEM

OUR PROPOSAL (I)

For each tweeters *i*, the distribution function of C_i (the sum of the n_i retweets) , that is $F_i(x) = P(C_i \leq x)$, can be found by means of a convolution between the distributions of n_i and x_i :

$$F_i(x) = \sum_{n_i=1}^{\infty} p(n_i) k^{n_i *}(x_i)$$

Where $k^{n_i *}$ indicates the n_i-fold convolution operator of the distribution k(.) with itself and, for each tweeters $p(n_i)$ is the distribution of the number of produced tweets and $k(x_i)$ is the distribution of the retweets.

EUROPEAN CENTRAL BANK

EUROSYSTEM

OUR PROPOSAL (II)

To complete the proposed model we need to specify two parametric distributions, one for the tweet production mechanism and one for the retweet patterns.

For the tweet production mechanism we propose and compare

- **Uniform discrete;**
- **Poisson:**

For the retweet patterns:

- **Poisson;**
- **Zipf-Mandelbrot (with the finite and infinite support)**
- **Negative Binomial**

All the possible combinations have been evaluated according to the Chi-square test and the Uniform-Negative Binomial results to be the best convolution.

EUROPEAN CENTRAL BANK

EUROSYSTEM

APPLICATION TO FINANCIAL TIMES TOP 2013 TWEETERS (I)

ID	h index	ID	h index	ID	h index	ID	h index
@abnormalreturns	16	@ECONOMISTHULK	48	@justinwolfers	48	@Queen_Europe	131
@AdamPosen	5	@economistmeg	6	@kathylienfx	10	@RedDogT3Live	9
@alaidi	10	@EpicureanDeal	9	@katie_martin_FX	10	@ReformedBroker	24
@Alea	2	@EU_Eurostat	31	@KavanaghKillik	5	@reinman_mt	3
@alexmasterley	18	@EU_Markt	10	@KeithMcCullough	7	@RencapMan	9
@andrealeadsom	7	@ezraklein	62	@LaMonicaBuzz	4	@RichardJMurphy	21
@andrewrsorkin	23	@FaullJonathan	5	@LaurenLaCapra	3	@ritholtz	12
@AngryArb	2	@felixsalmon	21	@Lavorgnanomics	15	@robertjgardner	2
@Austan_Goolsbee	3	@FGoria	9	@lemasabachthani	12	@SallieKrawcheck	10
@BergenCapital	22	@finansakrobat	5	@LorcanRK	5	@Scaramucci	8
@bespokeinvest	20	@firoozye	3	@mark_dow	9	@ScottMinerd	13
@bill_easterly	17	@footnoted	7	@MarkMobius	25	@SEK_bonds	7
@bobivry	4	@Fullcarry	3	@MatinaStevis	10	@SharonBowlesMEP	9
@bondvigilantes	20	@GCGodfrey	17	@mattyglesias	27	@Simon_Nixon	5
@BrendaKelly_IG	3	@greg_ip	14	@MBarnierEU	24	@SimoneFoxman	3
@BritishInsurers	7	@GSElevator	505	@michaelhewson	0	@SonyKapoor	24
@chrisadamsmkts	7	@GTCost	5	@MorrisseyHelena	7	@stlouisfed	21
@counterparties	8	@gusbaratta	4	@mtaibbi	113	@TedTobiasonDB	5
@CVecchioFX	6	@harmongreg	5	@NicTrades	4	@TheBubbleBubble	17
@DanielAlpert	8	@hblodget	34	@Nouriel	28	@TheNickLeeson	6
@danprimack	11	@hmtreasury	21	@OpenEurope	7	@TradeDesk_Steve	12
@DavidJones_IG	0	@howardlindzon	7	@Pawelmorski	12	@Trader_Dante	7
@davidmwessel	18	@Hugodixon	10	@pdacosta	18	@truemagic68	11
@DCBorthwick	4	@IvanTheK	11	@pensionlawyeruk	3	@WhelanKarl	10
@DKThomp	19	@jdportes	7	@PensionsMonkey	5	@WilliamsonChris	12
@Dorte_Hoppner	3	@Joe_Trading	2	@petenajarian	12	@WillJAitken	3
@DougKass	10	@JohnKayFT	14	@PIMCO	102	@World_First	6
@dsquareddigest	7	@JohnMannMP	13	@PIRCpress	4	@zerohedge	46
						@ZorTrades	6

EUROPEAN CENTRAL BANK

EUROSYSTEM

APPLICATION TO FINANCIAL TIMES TOP 2013 TWEETERS (II)

Using confidence intervals based on the stochastic *h* index to compare influential tweeters (2° through 5°)

ID	U-NB
@mtaibbi (observed h=113)	[109; 124]
@PIMCO (observed h=102)	[101; 112]
@ECONOMISTHULK (observed h=48)	[50; 62]
@justinwolfers (observed h=48)	[48; 59]

EUROPEAN CENTRAL BANK

EUROSYSTEM

APPLICATION TO ITALIAN FINANCIAL TWEETERS SOURCES AND SEMANTICS

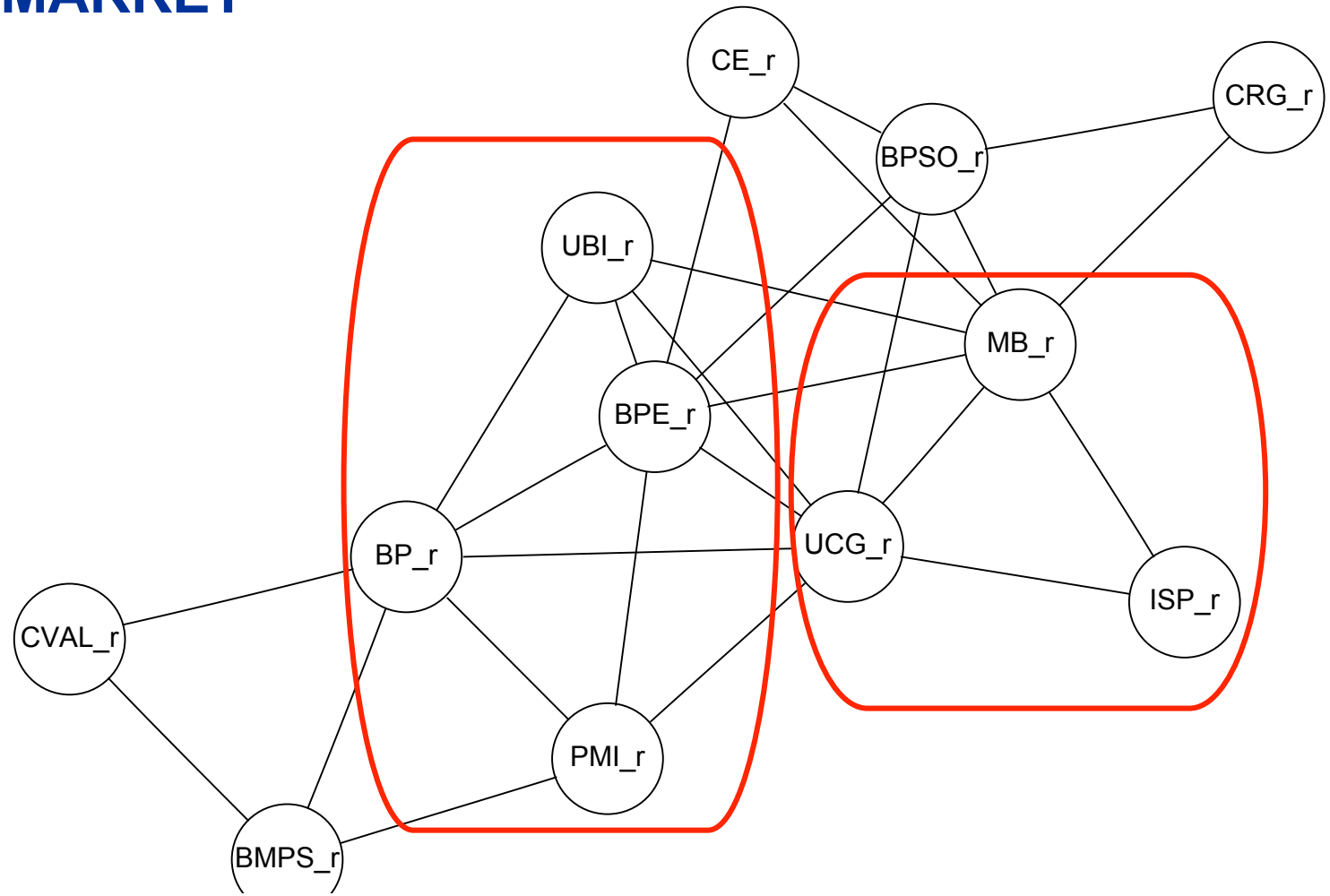
Source	H index
Sole 24 ore	18
Milano Finanza	7
Italia Oggi	6
Ansa Economia	4
Reuters Italia	8
Lavoce	24
Dagospia	14
Linkiesta	12

Sentiment Scale	Frequency
Very Bad (1)	0.90 %
Bad (2)	40.97%
Neutral (3)	0.44%
Good (4)	56.69%
Very Good (5)	1.01%

EUROPEAN CENTRAL BANK

EUROSYSTEM

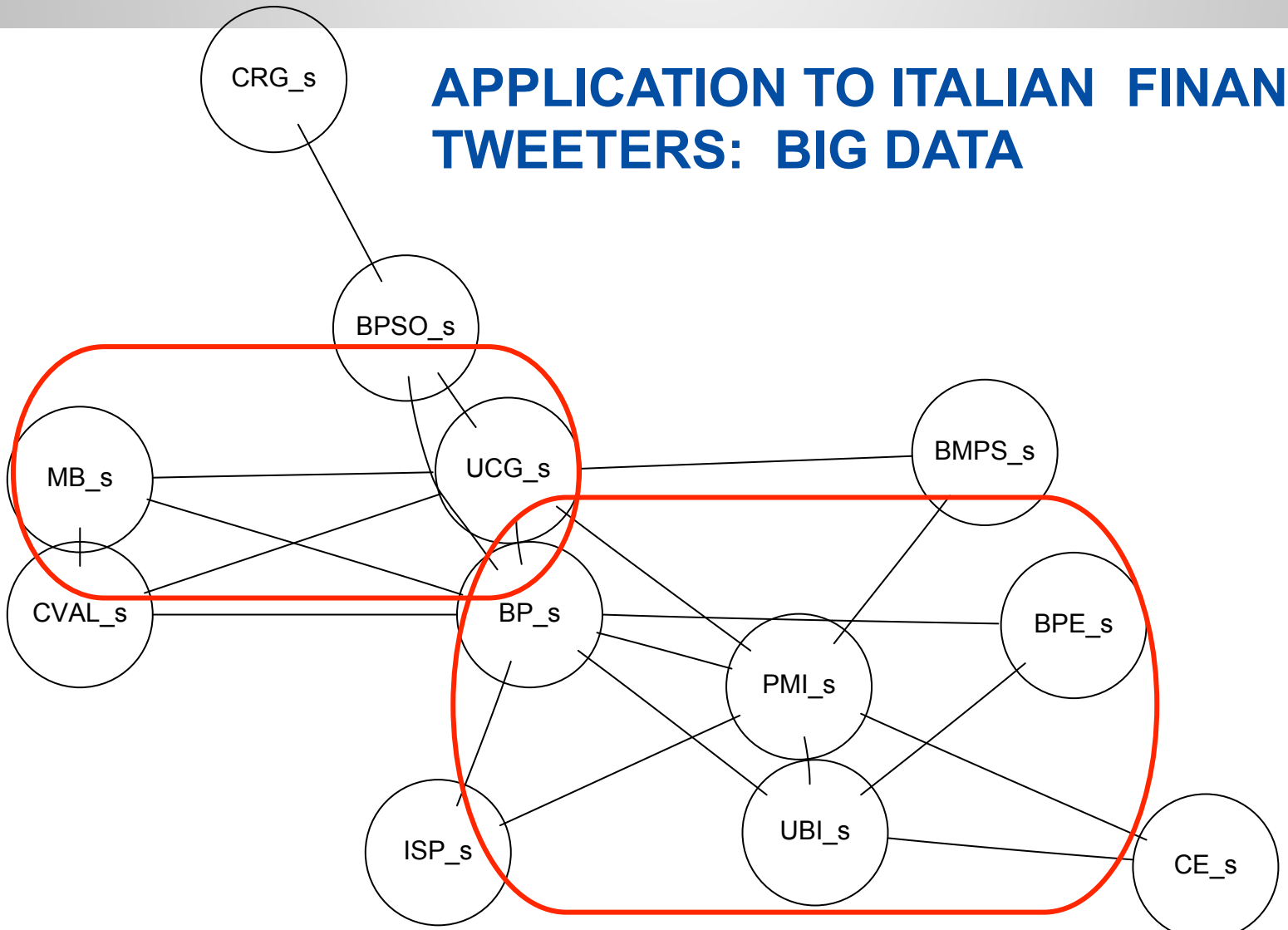
APPLICATION TO ITALIAN FINANCIAL TWEETERS: THE MARKET



EUROPEAN CENTRAL BANK

EUROSYSTEM

APPLICATION TO ITALIAN FINANCIAL TWEETERS: BIG DATA



*Analysis in collaboration with the semantic data analysis company EXPERT SYSTEMS