Attention to Bitcoin

Amirhossein Sadoghi

ESC Rennes School of Business, France

Modelling with Big Data and Machine Learning:
Interpretability and Model Uncertainty
The Bank of England and King's College London

4-5 November 2019

Outline

1. **Introduction**
   - Motivation
2. **Methodology**
   - Text-mining
   - Latent Dirichlet Allocation (LDA)
   - Measure of Information
   - Discontinuity Regression Design
3. **Data**
4. **Empirical Design**
5. **Empirical Results**
   - Results of LDA
   - Results for Bubble Phase
6. **Robustness Analysis**
7. **Conclusion**

Motivation

- The **resilience** of bubbles in the markets stems from the **attention** of market participants;
- The attentions might be governed by **media**;
- The **limited attention** or **(selective) inattention** of investors lead to **rational arbitrageurs** stay in the market;
- **Rational arbitrageurs** may know that a market will finally break down;
- In this limited-attention environment, the media which **convey specific information**
  - I might **attract** the attention of rational arbitrageurs
  - II **facilitate synchronized** decisions on **entering** or **exiting** such a market.

Research Question

How media exposure with regard to its intrinsic informational content can influence investors to enter or exit the market?

Summary of Project

- We empirically address above question regarding the **causes** of the **growth** and **bursting** of the **Bitcoin bubble**;
- We measure **attention** to news media characterised by **entropy** or **unusualness** [Glasserman and Mamaysky(2017)];
- We **classify** news items on Bitcoin into a group of subjects, then we investigate the effects of each group;
- We develop a new framework based on a set of **Regression Discontinuity (RD) experiments** associated with attention to news;
- We specify three **exogenous cut-off** points: **around the bubble event** as well as **before** and **after** the event;
- We then examine how these **informational flows** behave around the known cut-off points;

Literature Review

Our work is related to several bodies of literature on finance, machine learning:

- Literature on implications of attention for stock prices and volume

  [Hirshleifer et al.(2009)Hirshleifer, Lim and Teoh, Menzly and Ozbas(2010), Cohen and Frazzini(2008), Odean(1999), Odean(1998), Engelberg and Gao(2011)]

- Several theoretical models on the understanding of the underline mechanism of the asset bubble

  [Giglio et al.(2016)Giglio, Maggiori and Stroebel, Moinas and Pouget(2013), Lei et al.(2001)Lei, Noussair and Plott], Rational bubble theory:

  [Diba and Grossman(1988)]

- LDA model in economic studies:

  [Budak et al.(2014)Budak, Goel, Rao and Zervas, Nimark et al.(2016)Nimark, Pitschner et al., Bandiera et al.(2017)Bandiera, Hansen, Prat and Sadun, Mueller and Rauh(2018)].

Text mining

Text Mining is the **large-scale**, **automated processing** of plain text language in digital form to extract data that is converted into useful quantitative or qualitative information.

Soft and Hard Information in Finance

- Growing amount of financial data makes it more and more important to learn how to discover valuable information for financial decision making.
- In finance, there are typically two kinds of information:
  - **Soft information**: text, including opinions, ideas, and market commentary.
  - **Hard information:** numerical values, such as financial measures and historical prices.
- Text mining in finance aims to exploit soft information for financial models.

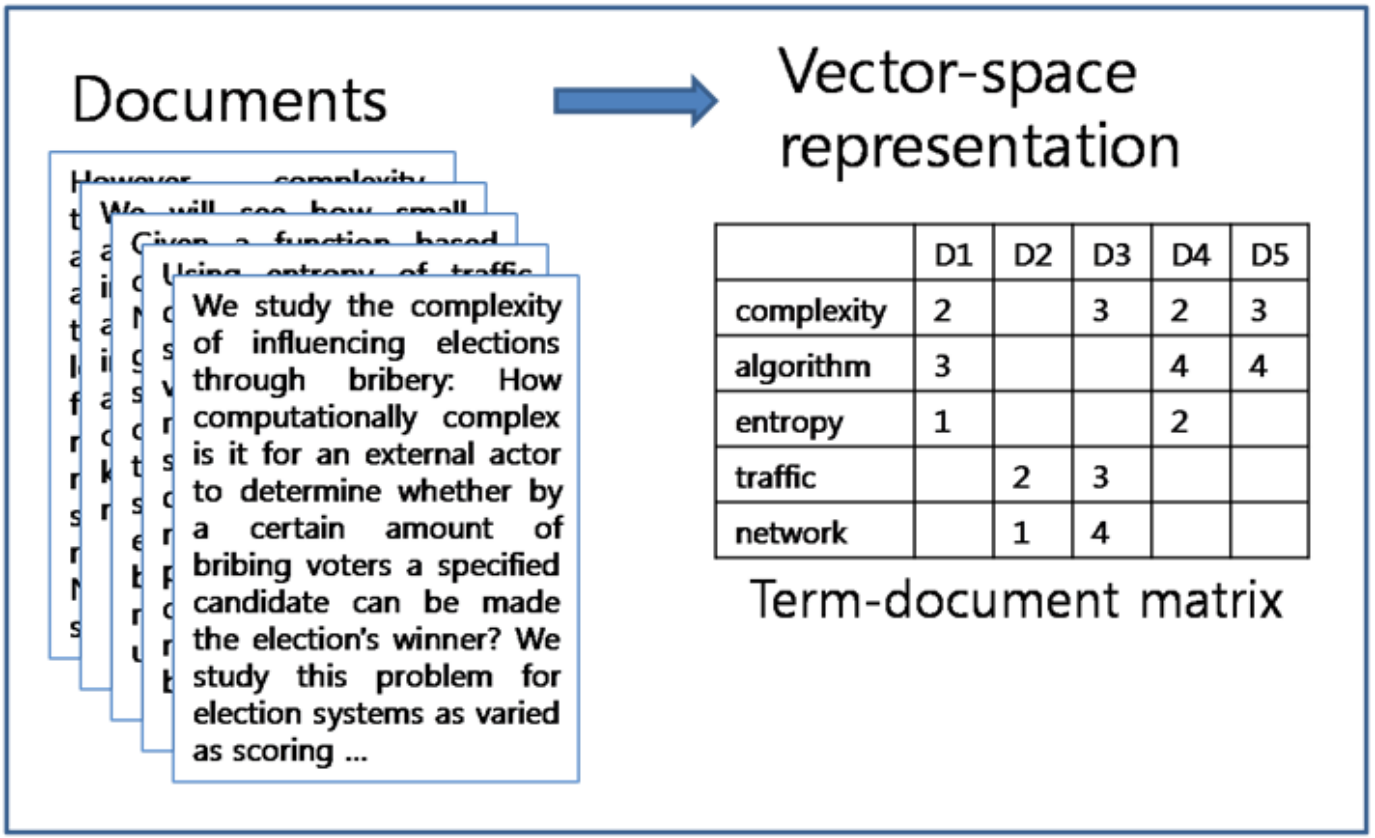
Preprocessing

Preprocessing Steps:

- Texts must be **tokenized** into smaller, more specific text features, such as words or word combinations.
- Removing "**stopwords**" : words designated in advance to be of no interest, and which are therefore discarded prior to analysis.
- Creating a **document-term matrix**.

Document-term Matrix (DTM)

The document term matrix (DTM) is one of the most common formats for representing a text corpus(i.e. a collection of texts) in a bag-of-words format.



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Topic Mining and Analysis: Motivation

- **Topic**: main idea discussed in text data or theme/subject of a discussion or conversation
  Different Granularities (e.g., topic of a sentence, an article, etc.)
- **Topic Modelling** (e.g.,[Blei et al.(2003)Blei, Ng and Jordan])
  Extract topics from text data and reveal their patterns;
  No consideration of time and structure of data;
- **Analysis of Causal Topics**;
- Many Applications in finance or economics require discovery of topics in text [Budak et al.(2014)Budak, Goel, Rao and Zervas, Nimark et al.(2016)Nimark, Pitschner et al., Bandiera et al.(2017)Bandiera, Hansen, Prat and Sadun, Mueller and Rauh(2018)].

LDA model

The LDA model in a hierarchical format:

- topics are distributed over documents,
- terms are distributed over topics,
- terms are distributed in documents.

$$P(\varphi, \theta, z, w) = P(\theta)P(\varphi) \prod_{d=1}^{D} P(z|\theta) \prod_{n=1}^{N} P(w|\varphi, z),$$

$\theta_{d(1:D)}$ is the topic distribution of document d,
$\varphi_{k(1:K)}$ is the term distribution of topic k,
$z_{dn(1:D,1:N)}$ is the topic distribution of the term n in document d,
$w_{dn(1:D,1:N)}$ is the distribution of term n in document d.

Measure of Information of Attention

Measuring the entropy of the estimator of the link between topic k and document d with parameter θ [Glasserman and Mamaysky(2017)].

$$H(\theta_{kd}) = -P(\theta_{kd}) \cdot log^{-1}(P(\theta_{kd})).$$

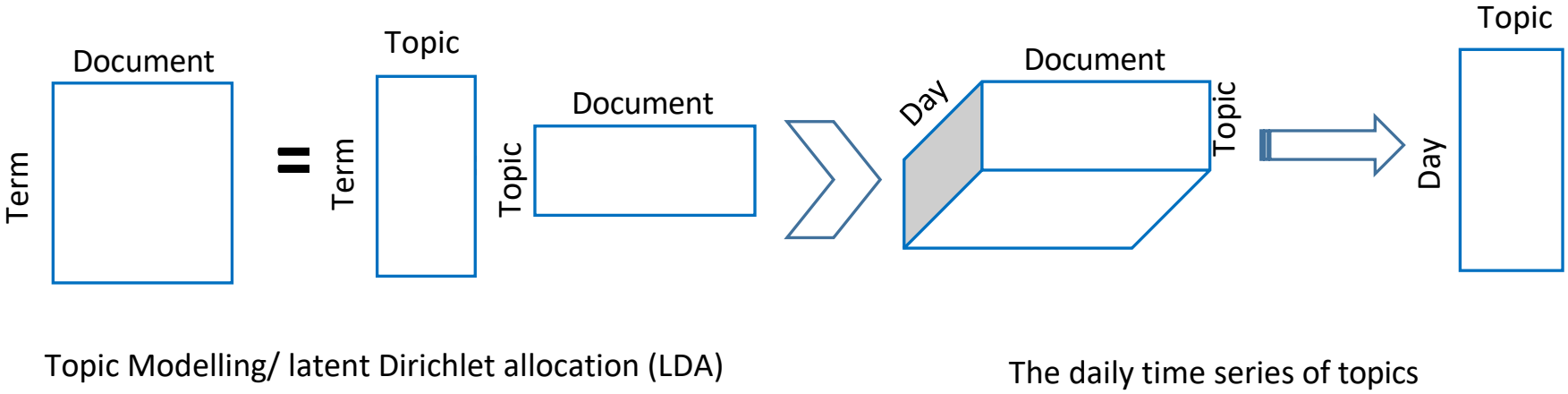
The daily value of intra-day entropy:

$$\bar{H}(\theta_{kt}) = \frac{\sum_{i=1}^{N_t} H(\theta_{ki})}{N_t},$$

N_t: the number of documents in day t

The Procedure of Creating Time Series of Topics

Figure: The Procedure of Creating Time Series of Topics



This figure presents the procedure of creating a daily time series of topics. The LDA method uses the Term-Document matrix and extracts the given number of topics. The later step aggregates the entropy of topics on a daily base and creates a time series of topics.

Discontinuity Regression Design

The Regression Discontinuity Design (RDD)

is a quasi-experimental technique where the assignment of the treatment and control is not random.

- Causal Relationship (LATE: the local average treatment effect)
- Forcing variable

  an assignment rule that one can use to assign individuals into treatment.
- Known cut-off

  as a function of one or more continuous variables that generates a discontinuity in the treatment assignment.
- Validity Tests

  No jump in outcome before treatment,
  No jumps in relevant covariates at cut off,
  No observations cannot be manipulate near the cutoff.
- Exclude all news about Price

Regression Discontinuity (RD)

A simple model of RD design with single measured feature:

$$y = a_0 + Df(x) + u,$$

The dependent variable y at around the cut-off point:

$$y^- \equiv \lim_{x \to \widetilde{x}^-} E[y|x] \quad \text{and} \quad y^+ \equiv \lim_{x \to \widetilde{x}^+} E[y|x].$$

$$\begin{aligned}
\Delta(\varepsilon) &= E\left[y|x = \widetilde{x}^-\right] - E\left[y|x = \widetilde{x}^+\right]. \\
&= f\left(\widetilde{x}^-\right) E\left[D|x = \widetilde{x}^-\right] + f\left(\widetilde{x}^-\right) E\left[D|x = \widetilde{x}^+\right] \\
&= f\left(\widetilde{x}^-\right) p\left(\widetilde{x}^-\right) + f\left(\widetilde{x}^-\right) p\left(\widetilde{x}^+\right).
\end{aligned}$$

Regression Discontinuity (RD) Continue

$$\begin{aligned}\lim_{\varepsilon\to 0}\Delta(\varepsilon) &= \lim_{\varepsilon\to 0} f\left(\widetilde{x}^{-}\right)p\left(\widetilde{x}^{-}\right)+f\left(\widetilde{x}^{-}\right)p\left(\widetilde{x}^{+}\right)\\ &= f\left(\widetilde{x}\right)\left(p^{-}-p^{+}\right).\end{aligned}$$

It leads to equation:

$$f\left(\widetilde{x}\right)=\frac{y^{-}-y^{+}}{p^{-}-p^{+}}.$$

- Fuzzy design:
  $p^{-}-p^{+}\neq 1,$
- Sharp RD design:
  $p^{-}-p^{+}=1.$

Data

- Data sample includes 244,703 online news articles (Jan 2014 to Dec 2018)
- The news dataset includes 11,650 web publications, 22,679 news wires, and 208,217 newsletters.
- Exclude all news about Price

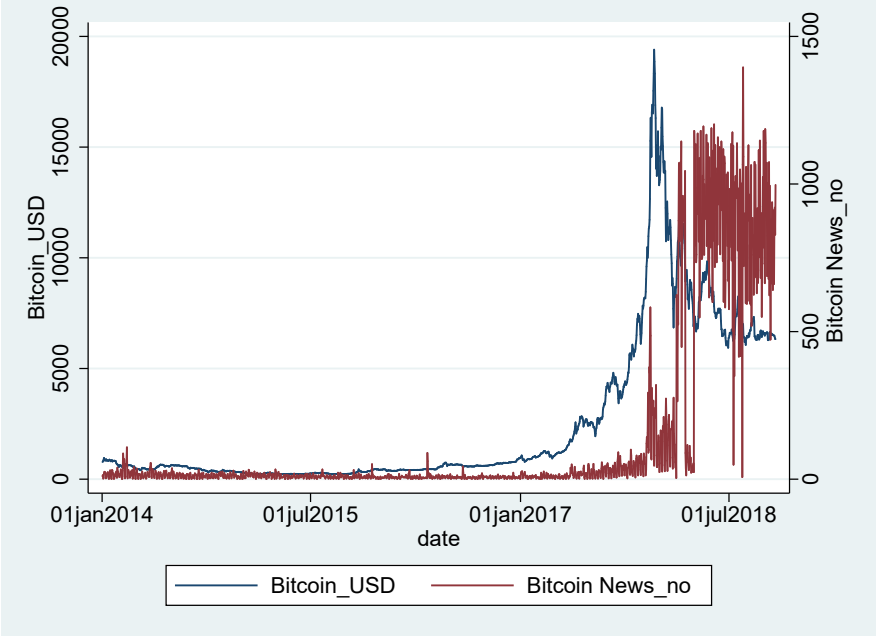


Figure: *Number of News Articles and Price of Bitcoin*

Empirical Design

Designing several experiments to analyse the effects of attention to Bitcoin in the news media.

- The running variables are taken from textual data and are exogenous.
- Using two days of lagged information on the entropy of news topics as running variables
- Cut-off points are determined independently of the running
  - Bitcoin price hit an all-time high below $20,000 (Dec 18, 2017)
  - Bitcoin price reached $1,000 for the first time (Jan 3, 2017).
  - Bitcoin price dropped to below $6000.
- We apply both fuzzy and sharp RD designs.

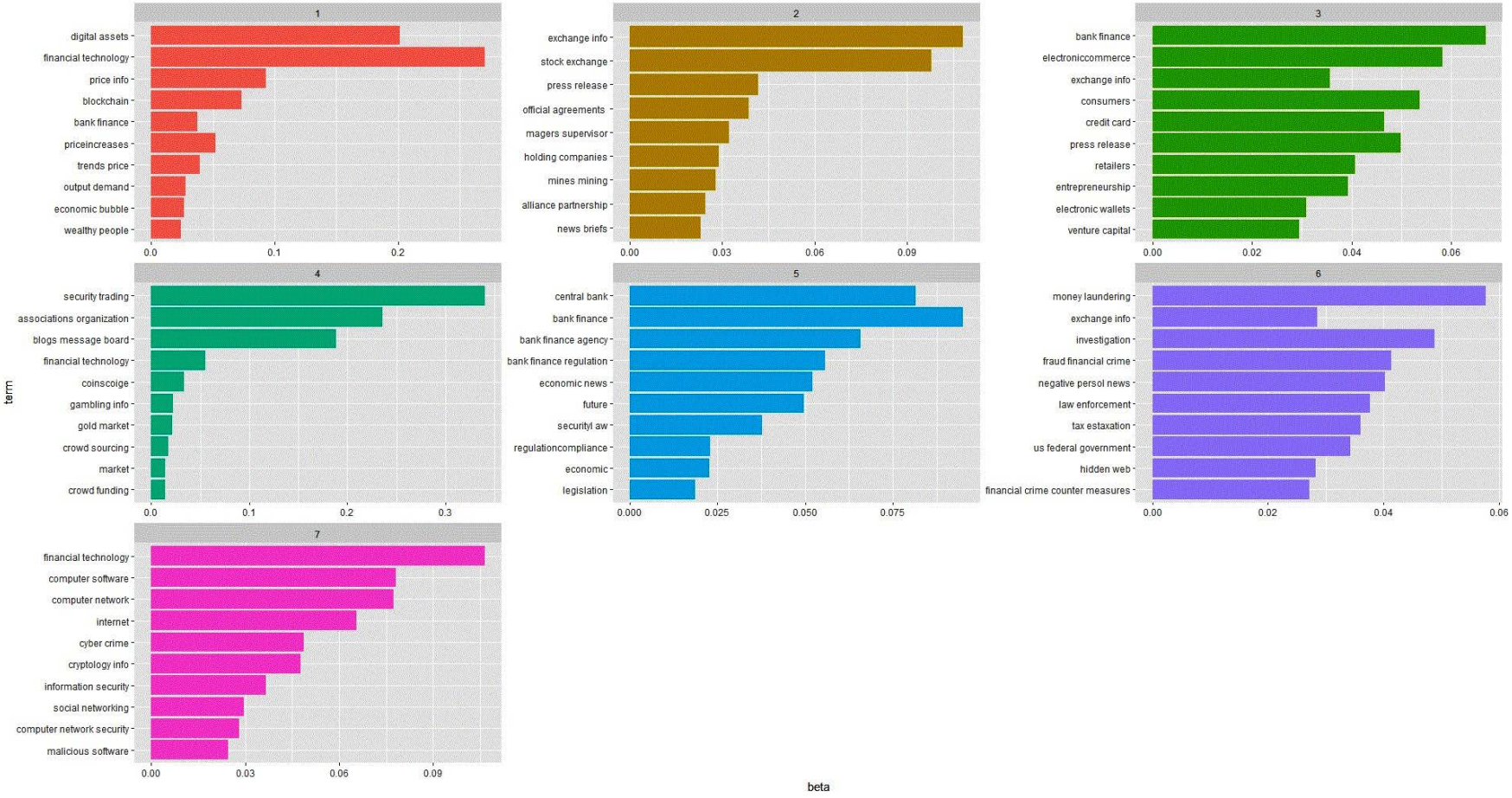
LDA Results

Topic 1: Fintech & Blockchain		Topic 2: Stock Exchange		Topic 3: Digital Banking-Finance		Topic 4: Security Trading	
Term	φ	Term	φ	Term	φ	Term	φ
financial tech	0.27	exchange info	0.108	bank finance	0.067	security trading	0.34
digital assets	0.201	stock exchange	0.07	electroniccomme	0.058	associa orga	0.236
decentralize	0.093	press release	0.042	consumers	0.054	blogs message	0.189
blockchain	0.073	official agreements	0.038	press release	0.05	financial technology	0.055
token-based	0.052	magers supervisor	0.032	credit card	0.047	coinscoige	0.033
market trend	0.039	holding companies	0.029	retailers	0.041	gambling info	0.022
bank finance	0.037	mines mining	0.028	entrepreneurship	0.039	gold market	0.022
output demand	0.028	stock exchange	0.028	exchange info	0.036	crowd sourcing	0.017
economic bubble	0.027	alliance partnership	0.024	electronic wallets	0.031	market	0.015
wealthy people	0.024	news briefs	0.023	venture capital	0.029	crowd funding	0.014
interviews	0.013	share holders	0.022	mobile application	0.028	gambling info	0.012
market changes	0.012	company profits	0.02	mobile application	0.023	market open close	0.011
industrial analysts	0.011	stockprice	0.02	internet ralted	0.022	time information	0.008
hedge fund	0.008	mines mining	0.018	electronic billing	0.021	government info	0.006
market size	0.008	company strategy	0.016	new products	0.019	voter voting	0.005
online trading	0.008	mine operations	0.016	atm	0.017	gambling info	0.003
intertiol trade	0.007	board directors	0.014	illegal drugs	0.017	productenhance	0.003
market research	0.007	company earnings	0.014	electronic bank	0.017	entertainments	0.002
market analysis	0.007	computer trading	0.013	conference	0.016	electroniccomm	0.001
research report	0.007	mine planning	0.013	debit card	0.015	profe associations	0.001

LDA Results

Topic 5 Central Bank & Regulation		Topic 6 Law & Tax Investigation		Topic 7 Computer Fintech	
Term	φ	Term	φ	Term	φ
bank finance	0.095	money laundering	0.058	financial technology	0.106
central bank	0.081	investigation	0.049	computer software	0.078
bank finance agency	0.066	fraud financial crime	0.041	computer network	0.077
bank finance regulation	0.056	negative news	0.04	internet	0.066
economic news	0.052	law enforcement	0.038	cyber crime	0.049
future	0.05	taxes taxation	0.036	cryptology info	0.048
security law	0.038	us federal government	0.034	information security	0.037
economic	0.023	exchange info	0.029	social networking	0.03
regulation compliance	0.023	hidden web	0.028	network security	0.028
legislation	0.019	financial crime counter measures	0.027	computer crime	0.024
talk meeting	0.018	litigation	0.027	malicious software	0.024
derivative instruments	0.017	special investigative forces	0.024	online security privacy	0.02
european union	0.017	arrests	0.023	digital signatures	0.019
interest rate	0.017	tax law	0.021	computer equipment	0.018
deflation	0.017	cyber crime	0.019	computer programming	0.016
public policy	0.017	law court tribuls	0.019	web programming	0.016
government info	0.016	lawyers	0.019	artificial intelligence	0.014
risk magement	0.015	terrorism	0.019	computing it	0.013
official approvals	0.014	controlled substance crime	0.018	network servers	0.013
us president candidate	0.014	criminal investigation	0.018	cloud computing	0.012

LDA Results


1
digital assets
financial technology
price info
blockchain
bank finance
priceincreases
trends price
output demand
economic bubble
wealthy people
0.0
0.1
0.2
2
exchange info
stock exchange
press release
official agreements
magers supervisor
holding companies
mines mining
alliance partnership
news briefs
0.00
0.03
0.06
0.09
3
bank finance
electroniccommerce
exchange info
consumers
credit card
press release
retailers
entrepreneurship
electronic wallets
venture capital
0.00
0.02
0.04
0.06
4
security trading
associations organization
blogs message board
financial technology
coinscoige
gambling info
gold market
crowd sourcing
market
crowd funding
0.0
0.1
0.2
0.3
5
central bank
bank finance
bank finance agency
bank finance regulation
economic news
future
securityl aw
regulationcompliance
economic
legislation
0.000
0.025
0.050
0.075
6
money laundering
exchange info
investigation
fraud financial crime
negative persol news
law enforcement
tax estaxation
us federal government
hidden web
financial crime counter measures
0.00
0.02
0.04
0.06
7
financial technology
computer software
computer network
internet
cyber crime
cryptology info
information security
social networking
computer network security
malicious software
0.00
0.03
0.06
0.09
term
beta

Falsification Test of RD Design: Bubble Phase

- Assumption 1:

  Assignment occurs through a known and measured deterministic decision rule,
- Assumption 2:

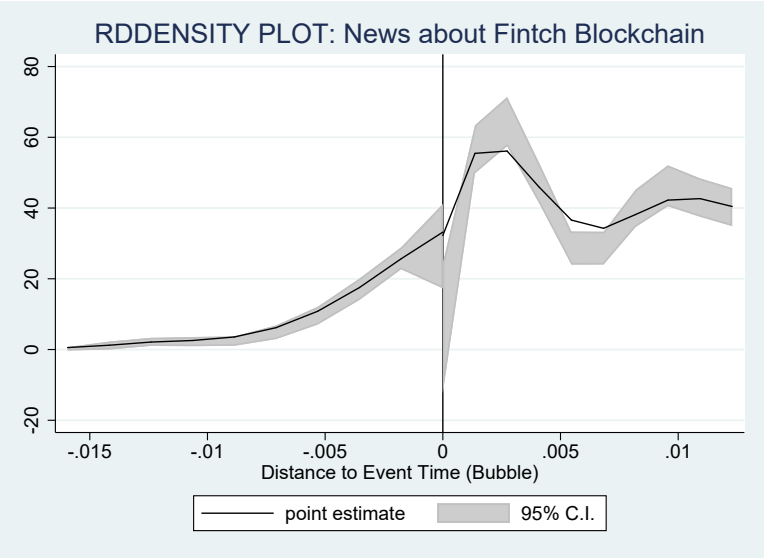
  Probability of assignment jumps at cut-off,
- Assumption 3:

  Local continuity.

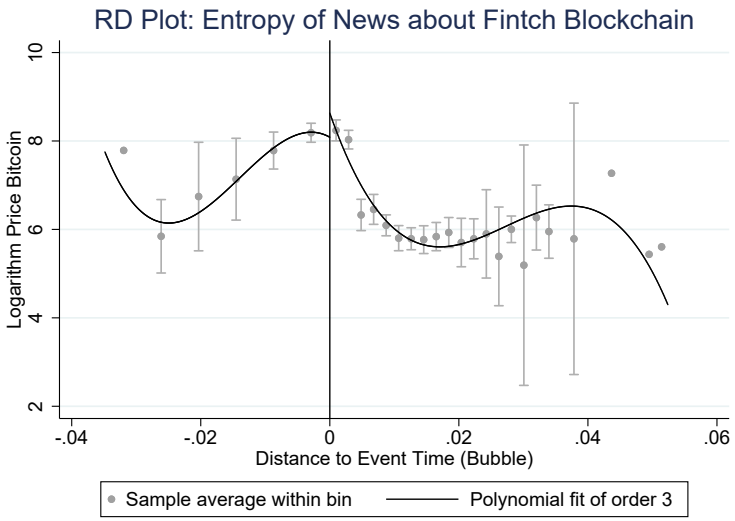
	Robust Effect	Robust P-value	Conventional Effect	Conventional P-value
Bank Finance Electronic	-14.779	0	37.735	.70342
Central Bank Reguation	21.973	.64856	20.933	.26165
Computer Fintech	111.113	.01016	92.544	.00024
Fintech blockchain	29.409	.03982	33.51	.86314
Law tax investigation	35.948	.15534	34.194	0
Security trading	63.872	.53353	65.39	.14893
Stock exchange	36.234	.12139	36.898	0

This table represents the results of manipulation tests by using local polynomial density functions. The null hypothesis of this test is that there is a discontinuity in running values. The P-value above 0.05 indicates the rejection of null of hypothesis.

Density Plot of Entropy of News Topics (Bubble Phase) Fintech Blockchain

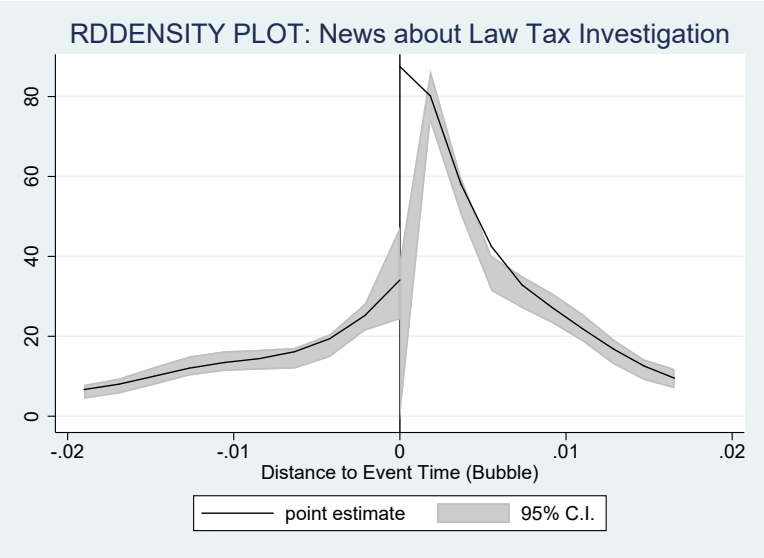


(a) Fintech Blockchain

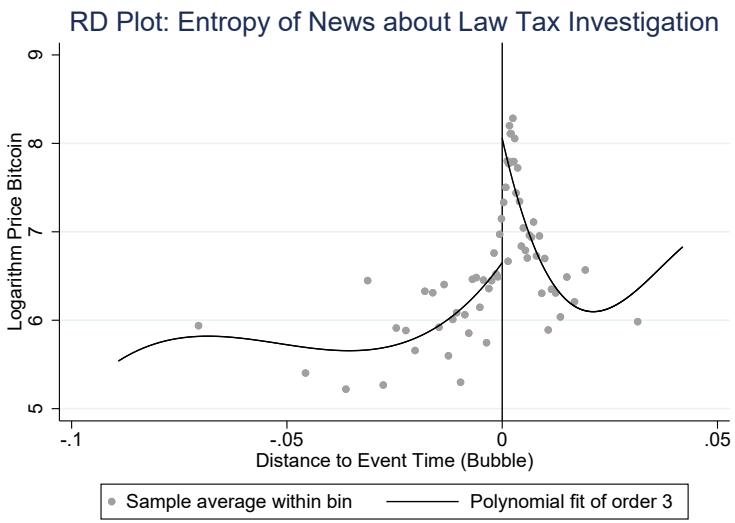


(b) Fintech Blockchain

Density Plot of Entropy of News Topics (Bubble Phase) Law Tax Investigation



(a) Law Tax Investigation



(b) Law Tax Investigation

RD Design: Attention to News Article about Bitcoin (Bubble Phase) Sharp

	(1)	(2)	(3)	(4)	(5)
Dependent Variable			(Ln.Price Bitcoin)		
Running Variable	CentralBank Regu	Fintech blockchain	Law tax investig	Security trading	Stock exchange
Sharp:					
Conventional	0.486**	0.485***	1.201***	0.177	0.0656
	(0.159)	(0.0740)	(0.147)	(0.135)	(0.139)
Bias-corrected	0.617***	1.011	1.311***	-0.0690	0.322*
	(0.159)	(0.154)	(0.147)	(0.135)	(0.139)
Robust	0.617**	1.011	1.311***	-0.0690	0.322
	(0.218)	(0.198)	(0.196)	(0.179)	(0.185)
Observations	1687	1687	1687	1687	1687

RD Design: Attention to News Article about Bitcoin (Bubble Phase) Fuzzy

	(1)	(2)	(3)	(4)	(5)
Dependent Variable			(Ln.Price Bitcoin)		
Running Variable	CentralBank Regu	Fintech blockchain	Law tax investig	Security trading	Stock exchange
Fuzzy:					
First Stage					
Conventional	$-.1211^{***}$	$.3065^{***}$	$.3703^{*}$	$-.0375^{***}$	$.1543^{*}$
	(.0547)	(.0768)	(.063)	(.035)	(.0632)
Robust	-.0269	$.2276^{***}$	.0776	-.051	-.0681
	(.0863)	(.1137)	(.0979)	(.0483)	(.0931)
Second Stage					
Conventional	2.979	6.982^{**}	2.622^{***}	7.585	0.718
	(2.472)	(4.616)	(0.574)	(7.233)	(1.409)
Bias-corrected	1.946	3.642	2.354^{***}	0.252	-1.245
	(2.472)	(2.408)	(0.574)	(7.233)	(1.409)
Robust	1.946	3.642	2.354^{**}	0.252	-1.245
	(3.694)	(3.796)	(0.887)	(10.26)	(2.049)
Observations	1687	1687	1687	1687	1687

Robustness Analysis

There are some issues about RD design which can reduce the validity and acceptability of the results.

- **Alternative Specifications:**
  Repeating the RD designs with different kernels functions and degrees of local polynomial.
- **Randomization:**
  "local randomization" [Lee and Card(2008)].
- **Misspecification of cut-off points**
  We shift two days the cut-off points to create tie-breaker experiments,

Conclusion

- We design several experiments to analyse the effects of attention to Bitcoin in the news media.
- The "**Causal Topics**" provide potential explanations for changes in markets,
- Findings:
  - Before the bubble phase:

    application in **Fintech** attracted the attention of investors to this unregulated market
  - During the bubble phase:

    uncertainty surrounding the **security** of this currency led investors to exit the market and resulted in the bursting of the bubble.
  - After the bubble phase:

    An ongoing discussion focused on **regulating** crypto-currency is the main factor shaping this market.
- **Future work**: Issues related to causality analysis ("global" and "local" causality) on Volatility of Bitcoin

References

Bandiera, O., Hansen, S., Prat, A., Sadun, R., 2017.
Ceo behavior and firm performance.
Technical Report. National Bureau of Economic Research.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003.
Latent dirichlet allocation.
Journal of machine Learning research 3, 993–1022.

Budak, C., Goel, S., Rao, J.M., Zervas, G., 2014.
Do-not-track and the economics of third-party advertising.
Boston University, School of Management Research Paper 2505643.

Cohen, L., Frazzini, A., 2008.
Economic links and predictable returns.
The Journal of Finance 63, 1977–2011.

Diba, B.T., Grossman, H.I., 1988.
The theory of rational bubbles in stock prices.
The Economic Journal 98, 746–754.

Engelberg, J., Gao, P., 2011.
In search of attention.
The Journal of Finance 66, 1461–1499.

Giglio, S., Maggiori, M., Stroebel, J., 2016.
No-bubble condition: Model-free tests in housing markets.
Econometrica 84, 1047–1091.

Glasserman, P., Mamaysky, H., 2017.
Does unusual news forecast market stress?
Technical Report. working paper.

Thank you for your attention!

2020 Computing: Everything everywhere

Declan Butler, nature, Vol 440, Issue no. 7083, 23 March 2006

In the medical sciences, researchers will be able to mine up-to-the-minute databases instead of painstakingly collecting their own data

The understanding of diseases, and the efficacy of treatments will be dissected by ceaselessly monitoring huge clinical populations

It will be a very different way of thinking, sifting through the data to find patterns.

A two-way street to science's future

Ian Foster, nature, Vol 440, Issue no. 7083, 23 March 2006

Science is increasingly about information: its collection, organization and transformation

George Djorgovski: "Applied computer science is now playing the role which mathematics did from the seventeenth through the twentieth centuries: providing an orderly, formal framework and exploratory apparatus for other sciences"

Science is becoming less reductionist and more integrative

LINKÖPINGS UNIVERSITET

Data mining and statistical learning:
lecture 1a