Dynamic Topic Modeling for Bitcoin Message Fora

Elisabeth Bommes
Cathy Yi-Hsuan Chen
Ernie Gin Swee Teo
Wolfgang K. Härdle
Marco Linton

Ladislaus von Bortkiewicz Chair of Statistics

Sim Kee Boon Institute for Financial Economics

International Research Training Group

Humboldt-Universität zu Berlin
lvb.wiwi.hu-berlin.de

www.case.hu-berlin.de

irtg1792.hu-berlin.de

Cryptocurrencies

- 1032 cryptos

- coinmarketcap.com

- Social media mining

All Currencies

#	Name	Symbol	Market Cap	Price	Circulating Supply	Volume (24h)	% 1h	% 24h	% 7d
1	Bitcoin	BTC	$28,385,989,541	$1738.52	16,327,675	$675,146,000	-2.09%	-2.88%	11.27%
2	Ethereum	ETH	$8,105,661,203	$88.52	91,573,871	$117,790,000	0.02%	-3.34%	-6.01%
3	Ripple	XRP	$7,922,190,998	$0.206814	38,305,873,865 *	$117,393,000	7.19%	2.43%	109.81%
4	Litecoin	LTC	$1,401,208,541	$27.43	51,084,557	$107,344,000	-0.64%	-7.26%	4.76%
5	NEM	XEM	$1,094,094,000	$0.121566	8,999,999,999 *	$10,714,900	-2.15%	-0.44%	69.93%
6	Dash	DASH	$621,511,724	$85.21	7,293,746	$15,712,300	-1.19%	-6.29%	-14.87%
7	Ethereum Classic	ETC	$558,737,086	$6.10	91,584,833	$13,966,400	-0.25%	-6.75%	-14.81%
8	Monero	XMR	$404,217,384	$27.97	14,454,351	$7,561,560	-0.37%	-6.39%	-2.22%
9	Stellar Lumens	XLM	$355,391,179	$0.037254	9,539,806,433 *	$49,138,900	2.53%	-12.83%	135.66%
10	Augur	REP	$179,304,400	$16.30	11,000,000 *	$1,376,920	-0.20%	-4.25%	-6.47%
11	Golem	GNT	$176,524,792	$0.215208	820,252,000 *	$3,469,760	0.51%	-7.70%	-12.85%
12	Steem	STEEM	$166,201,327	$0.712369	233,307,916	$3,889,570	-0.83%	-11.30%	104.50%
13	Round	ROUND	$163,167,700	$0.191962	850,000,000 *	$1,395,820	-0.34%	45.48%	176.10%
14	MaidSafeCoin	MAID	$138,089,128	$0.305134	452,552,412 *	$4,994,820	4.12%	-5.02%	5.20%
15	Dogecoin	DOGE	$128,429,259	$0.001174	109,355,476,924	$5,232,350	0.99%	-5.50%	36.02%
16	Stratis	STRAT	$123,851,359	$1.26	98,382,961 *	$1,490,670	0.62%	-2.78%	5.01%
17	Zcash	ZEC	$123,607,188	$95.59	1,293,081	$4,874,940	0.35%	-4.10%	-6.87%
18	Gnosis	GNO	$119,208,457	$107.92	1,104,590 *	$789,488	0.55%	-4.49%	2.46%

All ▾   Currencies ▾   Assets ▾   USD ▾   ← Back to Top 100

ORIX

CRIX

☑ Unlock the value of dark data to transform it into smart data

hu.berlin/crix

Current cryptos in CRIX

Current cryptocurrencies in CRIX

Coin	Name	Price (in $)	Market Cap (in $K)	Volume (in $K)
1	btc	998.05	16,045,986	3,470,830
2	eth	8.06	705,615	7,812
3	xrp	0.01	231,687	418
4	ltc	4.50	221,302	427,541
5	xmr	13.62	186,132	3,153
6	etc	1.40	122,421	3,568
7	dash	11.39	79,673	456
8	maid	0.10	43,675	455
9	rep	3.94	43,340	229
10	steem	0.16	37,154	117

Showing 1 to 10 of 25 rows   10 ▲   records per page   ‹ 1 2 3 ›

CRIX

Cryptos

⊡ Cryptocurrencies have become mainstream

⊡ Domain of technophiles and radicals ?

⊡ traded on many exchanges (blockchain)

⊡ bottomless source of information

⊡ Crowd wisdom a powerful indicator of major events

What can be done?

▣ Dynamic topic modelling (DTM), text mining and machine learning

▣ detect events: new trends in currencies, fraudulent schemes

▣ unsupervised DTM opinion evolution of topics

▣ test hypothesis of self-fulfilling prophecies and herding behaviour

▣ Smailovic et al. (2013) : improve prediction from Twitter sentiments

Introduction

⊡ Amount of unstructured dark data is growing

⊡ Conventional ways: read, search and links?

⊡ Theme based exploration offers more possibilities

Desiderata

⊡ label/ tag all documents with corresponding topics

⊡ „Zoom in" and „zoom out" to find specific or broader themes

⊡ Find new connections and changes over time

⊡ But to read all the available documents is not in human power

⊡ Solution: probabilistic topic modeling

CRIX

Word distribution

Topic: NASDAQ

Topic: Shakespeare
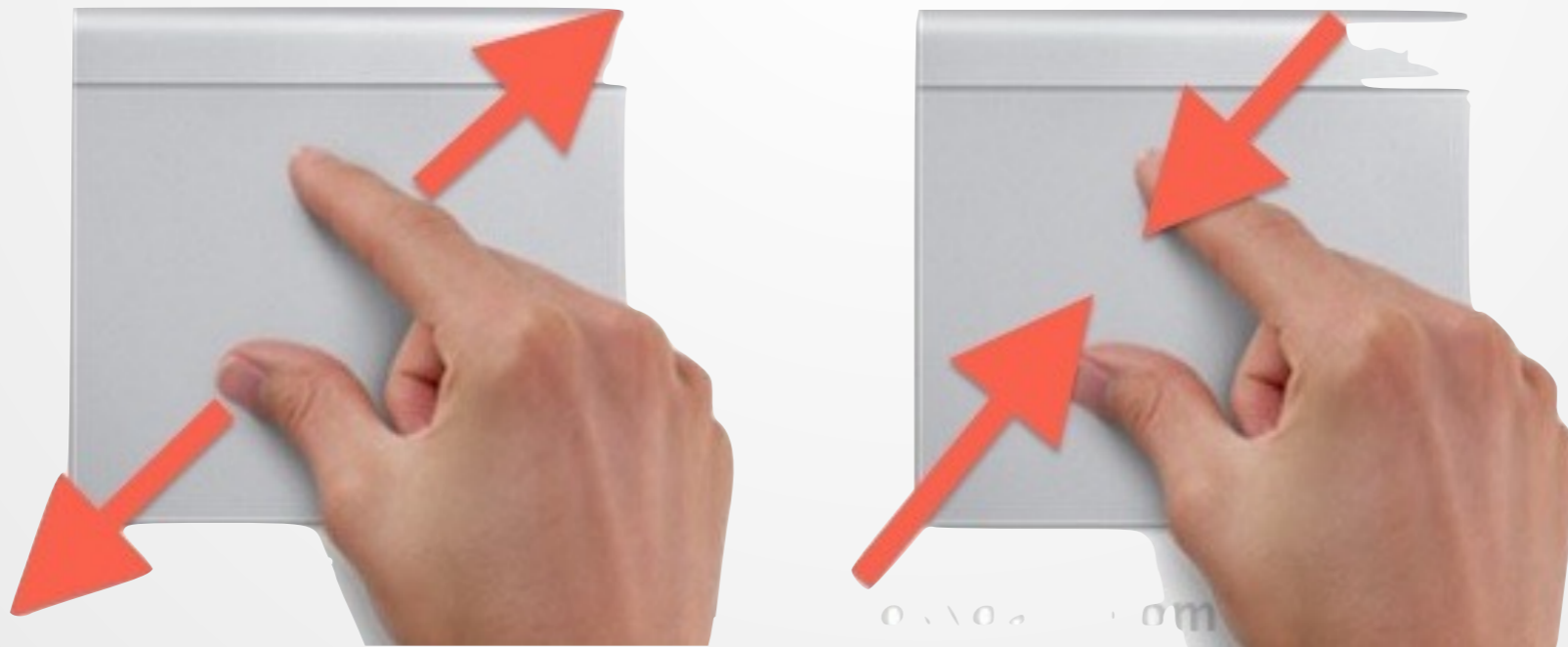
Latent Dirichlet allocation (LDA)

⊡ Simple topic model

⊡ Document may exhibit several topics

⊡ Each topic is a distribution over a fixed vocabulary

⊡ Topics are generated first, before the documents

⊡ Data is arising from a generative process

CRIX

LDA: Generative Process

Words are generated in a two-stage process:

1. Randomly choose a distribution over topics.

2. For each word in the document

   a) Randomly choose a topic from the distribution over topics

   b) Randomly choose a word from the corresponding distribution over the vocabulary.

Example 1

- Observed data

  - 2 documents

  - 155 distinct words

  ⊡? number of topics

Silent Night! Holy Night!

From the Third (unpublished) Part of "HYMNS AND MUSIC FOR THE YOUNG." By permission of the Author.

1. Si-lent night! Ho-ly night! All is calm, all is bright; Round yon Virgin Mother and Child!

Jingle Bells

Dashing through the snow
In a one horse open sleigh
O'er the fields we go
Laughing all the way
Bells on bob tails ring
Making spirits bright
What fun it is to laugh and sing
A sleighing song tonight

Oh, jingle bells, jingle bells
Jingle all the way
Oh, what fun it is to ride
In a one horse open sleigh
Jingle bells, jingle bells
Jingle all the way

- Unobserved variables

  - topic assignment of words

  - per-document topic proportions (final result)

Example 2

Example 2: Shakespeare's tragedies

$$T = \{art, bear, call, day, dead, dear, death, die, eye, fair, father, fear$$
$$friend, god, good, heart, heaven, king, ladi, lie, like, live, love,$$
$$make, man, mean, men, must, night, queen, think, time\}$$
$$= \{t_1, \ldots, t_{32}\}$$

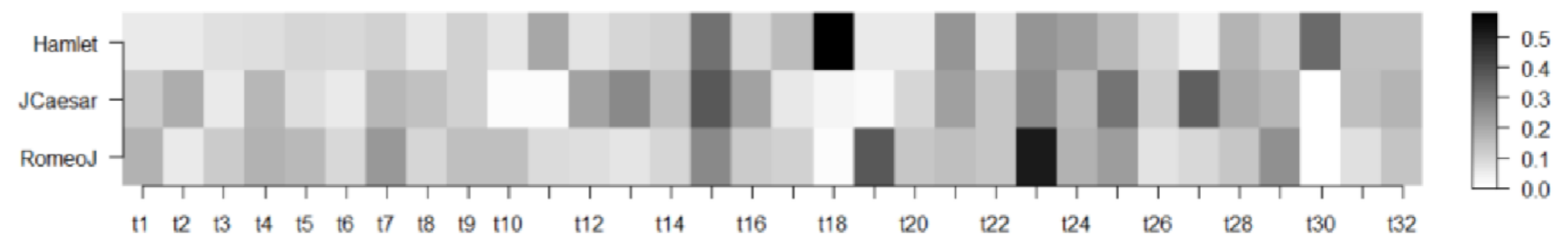T – special vocabulary selected among 100 most frequent words.

Figure 10: Heatmap of T in 3 tragedies

- ⊡ Doc 1: Hamlet (words 16769)

- ⊡ Doc 2: Julius Caesar (words 11003)

- ⊡ Doc 3: Romeo and Juliet (words 14237)

- ⊡ ? number of topics

Example 3

⊡ NASDAQ docs 2009 - 2017.1

⊡ Lemmatization, negation

⊡ ? number of topics
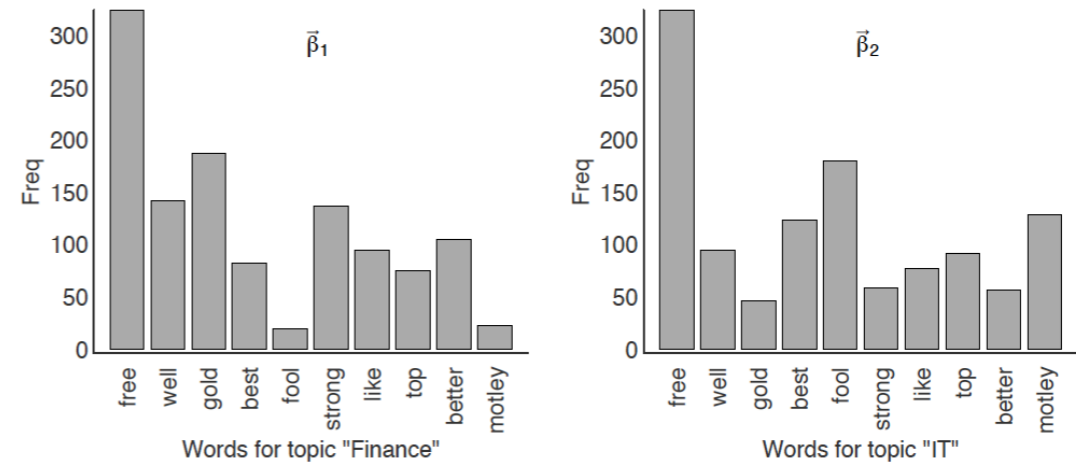
Figure 2: Distribution of words by topic ($\vec{\beta}_1$ and $\vec{\beta}_2$)

XFGdtmWDistr

Word	Freq. (in k)	Freq. for Top 5 Sectors
free	649	10
well	238	9
gold	235	1
best	207	9
fool	200	5
strong	196	5
like	172	5
top	167	3
better	162	0
motley	152	2

Table 1: Most frequent words used in NASDAQ articles

Example 4

☑ Number of topics K = 3, number of documents D = 1

Bayesian machinery

⊡ Main difference to classical (frequentist) statistics:

⊡ parameters are rv's and not fixed unknown quantities

⊡ combines prior information with data, within a decision theoretical framework

⊡ provides inferences that are conditional on the data and are exact,

⊡ obeys the likelihood principle

LDA OK for more than topicing Christmas songs?

▣ Dynamic Topic Modelling for Cryptocurrency Community Forums

▣ „an indicator for fraudulent schemes constructed using DTM, text mining and unsupervised machine learning"

▣ „study how opinions and the evolution of topics are connected with big events in the cryptocurrency universe"

Words are the new numbers

▣ Vegard H. Larseny and Leif Anders Thorsrud (Norges Bank) decompose a major business newspaper using LDA

▣ Simple hypothesis: the more intensive a given topic is represented the more likely it is that this topic is of importance for the economy's future

▣ Result: topics have predictive power for key economic variables and, especially noteworthy, for asset prices

CRIX

LDA and Topic detection
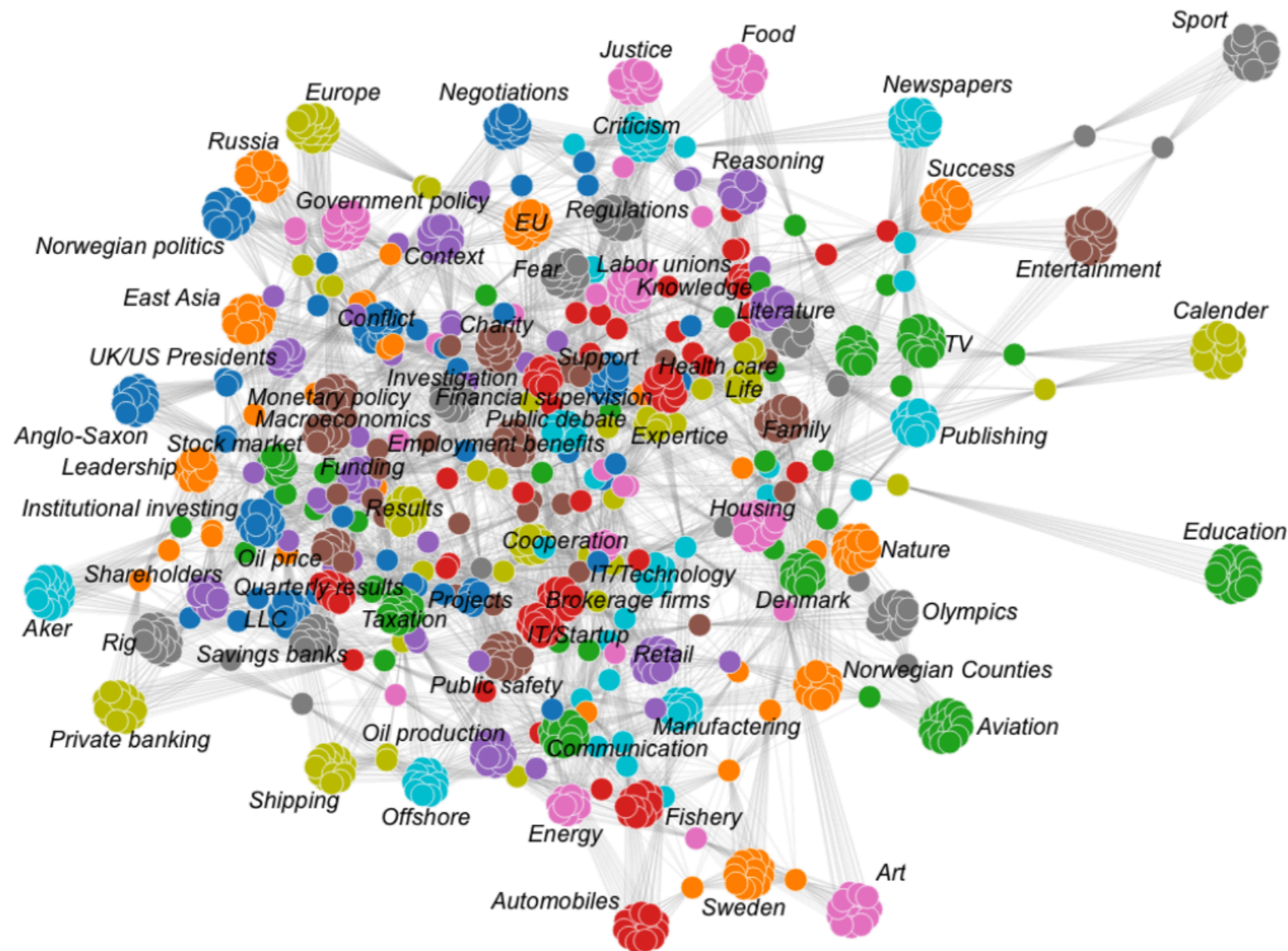
Figure 1. DN visualized using a topic nett. Different colors are used for words in different topics. Words that belong to the same topic have an edge between them.

New idea

⊡ Use the approach to measure inflation expectations (IE)

⊡ Do news have predictive power for IE?

⊡ Use text mining to detect news relevant for inflation expectations, perception of the central bank announcements by the media etc

⊡ Compare inflation expectation measure based on news with other measures such as T-bond differences etc

⊡ Same can be done for options, Chen, Härdle, Liu (2017)
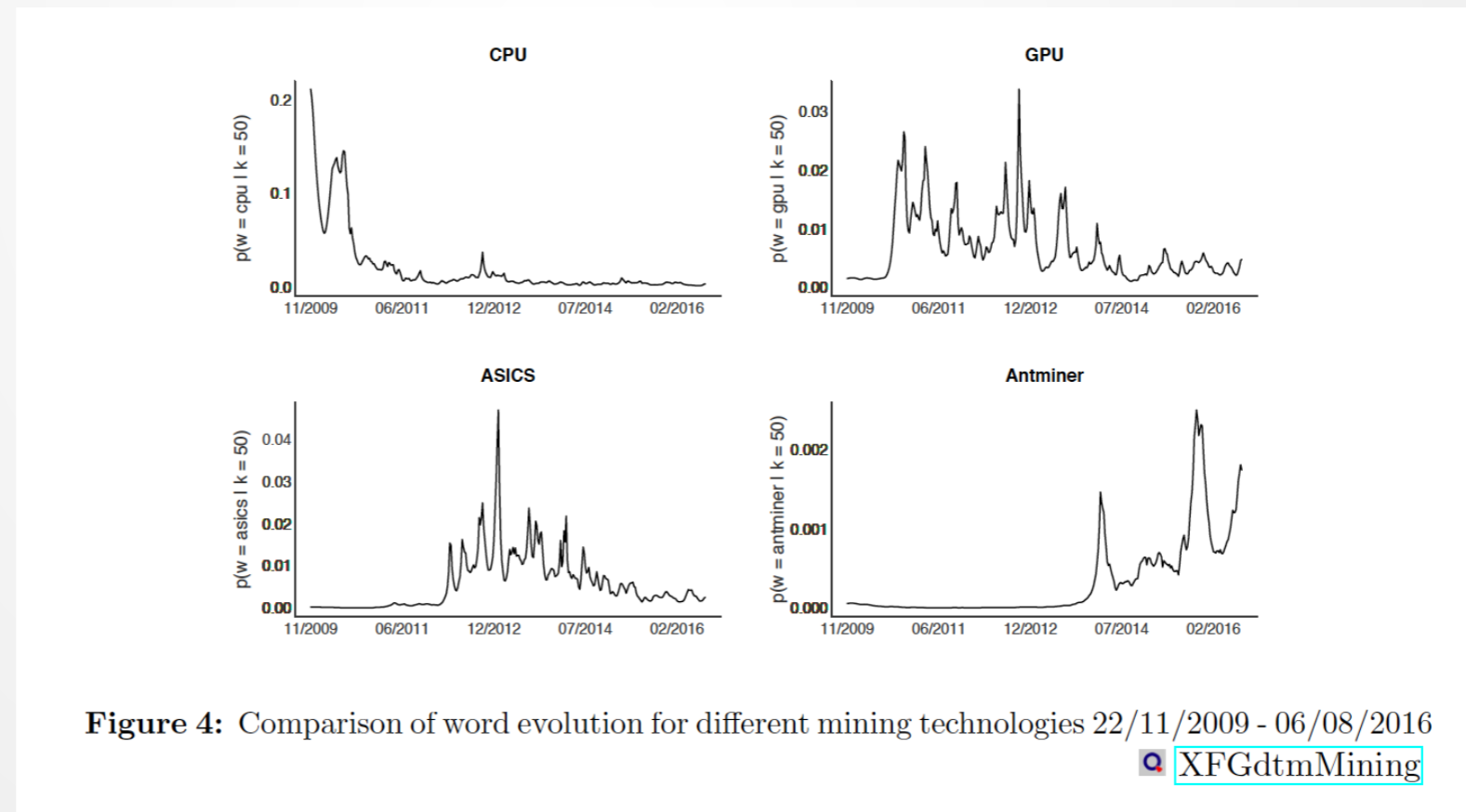
DTM for Bitcoin message boards

▣ Observe sub forums Economics, Bitcoin Discussion, Altcoin Discussion, Speculation, Scam Accusations

▣ Run with weekly data from 20091122 to 20160806 period

▣ Notable Topics:

Topic Number	Most Probable Words
1	value, gold, bar, dollar, rate, demand, interest, asset
2	business, casino, house, trust, gambling, run, strategy, player
5	government, control, criminal, law, study, regulation, state, rule
7	use, service, option, cash, good, spend, fiat, convert
12	account, payment, fund, card, paypal, party, merchant, credit
18	score, online, pay, shop, bill, product, purchase, phone
20	wallet, key, paper, computer, storage, code, data, secure
23	price, trade, market, trader, drop, volume, sell, stock
24	trading, term, hold, buy, pump, dump, earn, gamble
30	exchange, bitfinex, lesson, cryptocurrency, crash, platform, altcoins, popularity
32	investment, risk, invest, aim, impact, salary, making, way
33	year, altcoins, end, today, adoption, prediction, happen, trend
35	transaction, block, fee, chain, confirmation, hour, minute, hardfork
38	altcoin, company, loss, hack, scam, hacker, scammer, road
42	bank, system, security, fiat, banking, role, function, institution
45	ethereum, split, advantage, issue, side, change, fork, core
48	forum, post, topic, member, bitcointalk, thread, index, php
50	mining, miner, network, power, pool, cost, reward, electricity

Table 2: Notable topics from 50 topic model on Bitcoin Discussion subforum from 2016/07/31 to 2016/08/06

CRIX

DTM for Bitcoin message boards

- ▫ Bitcoins produced by CPU mining: worth less than cost to mine
- ▫ GPU mining came into play 2010
- ▫ ASIC application specific integrated circuits
- ▫ First ASICs („AvalonProject") in 2012, 2013.1 release of first chip
- ▫ Antminer, a brand of ASICS, current top of the line

Figure 4: Comparison of word evolution for different mining technologies 22/11/2009 - 06/08/2016

XFGdtmMining

DTM for Bitcoin message boards

⊡ **The event:** insolvency of the MtGox Bitcoin exchange in 2014. MtGox started

2007 trading Magic: The Gathering Online trading cards which is where it got its

name (Magic: The Gathering eXchange). In 2010, however, it was rebranded as

one of the first exchanges where people could buy and sell Bitcoins. The

exchange grew gradually and watched the price of Bitcoin go from less than USD

0.1 in 2010 to parity with the US dollar in 2011. At this point however, the owner

of MtGox decided to sell the exchange in order to dedicate himself to `other

projects'.

CRIX

DTM for Bitcoin message boards

⊡ An internal email dating back from after the sale of the exchange revealed that

already 80,000 Bitcoins (worth over $60,000) had already been missing before

any of the public fiascos had occurred. Three months later that a major event

occurred. 60,000 accounts were exposed publicly and a compromised MtGox

auditors account was used to create huge sell orders and crash the Bitcoin price

from $17.51 to $0.01. Many of the exposed accounts were used to steal coins

from other bitcoin services due to password reuse. However, unlike many other

Bitcoin services, MtGox managed to recover its reputation and became the

largest Bitcoin exchange, handling 70% of all trades worldwide.

DTM for Bitcoin message boards

⊡ Fast forwarding to 2013, when their real problems began, in June withdrawals of USD were suspended and even though a couple of weeks later in July it had been announced that withdrawals had fully resumed, as of September few withdrawals had successfully been completed. Complaints piled up over the next few months and on 7 February 2014 all Bitcoin withdrawals had been suspended for good. On the 24th of February all activities had halted, the website went offline and a leaked internal crisis management document claimed that 744,408 Bitcoins (worth almost half a billion dollars) had been lost and the company was insolvent.

CRIX

DTM for Bitcoin message boards

⊡ Topic 23 = Bitcoin trading

⊡ Topic 38 = scams and hacks

⊡ Topic 23: peak in 2013.6, transaction issues first occurred; trailing off when
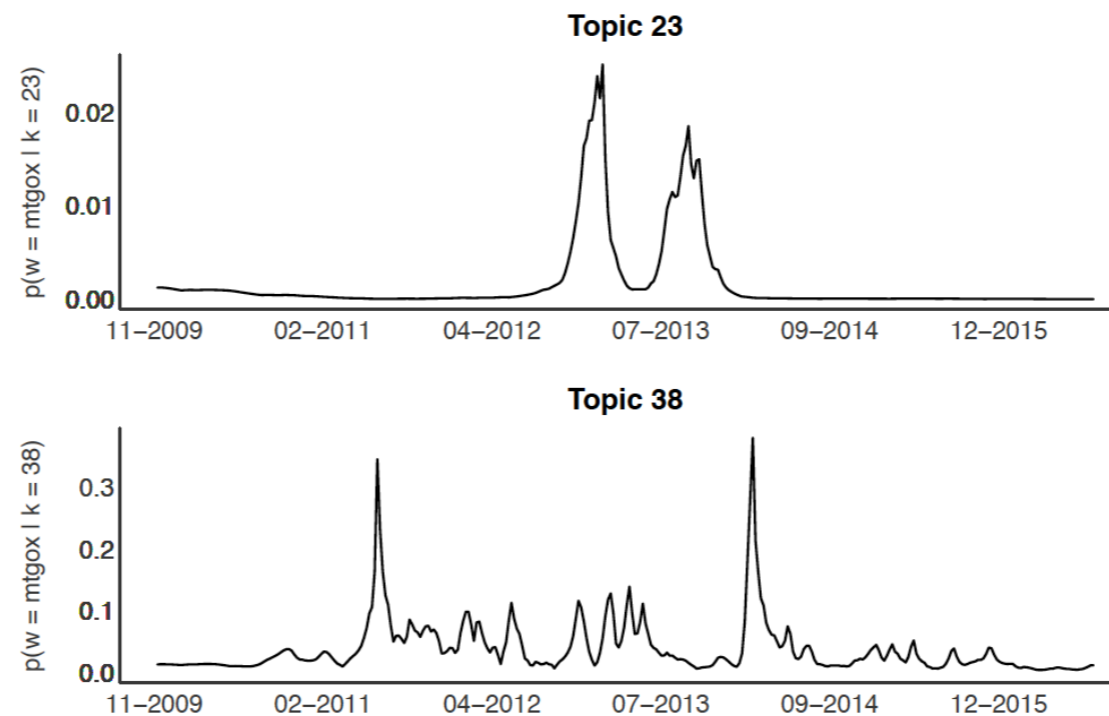
   MtGox starts to gain momentum in topic 38

XFGdtmMtGox

Topic 23

Topic 38

Figure 5: MtGox word evolution 22/11/2009 - 06/08/2016

XFGdtmMtGox

DTM for more events in the Bitcoin universe

⊡ Evaluate the effectiveness of topic models

⊡ list of Bitcoin services which have been victims of hacks:

   https://bitcointalk.org/index.php?topic=576337.0

⊡ basis for event discovery validation

DTM for more events in the Bitcoin universe

◦ List of 37 events on 33 different Bitcoin

  services. For each word determine which

  topics the word achieves a topic prominence

  larger than a low threshold.

◦ Topic prominence of the words conditioned on

  topics through time and determine a significant

  event.

Event	Dates	Topic
Ubitex* (1,138b)	2011-04 to 2011-07	None
Allinvain	2011-06-13	23
MtGox	2011-06-19	23
Mybitcoin	2011-06-20, 2011-07	23
Bitomat	2011-07-26	23
Mooncoin	2011-09-11	23
Bitscalper	2012-01 to 2012-03	23
Linode	2012-03-01	23
Betcoin* (3,171b)	2012-04-11	None
Bitcoinica	2012-04-12, 2012-07-13	23
Btc-e	2012-07-13	12
Kronos	2012-08	23
Bitcoin Savings and Trusts	2012-08-28	23
Bitfloor	2012-09-04	23
Btcguild* (1,254b)	2013-03-10	None
OkPay (main victim of 2013 Fork)	2013-03-11	30
Ziggap* (1,708b)	2013-02 to 2013-04	None
Just-Dice	2013-07-15	23
Basic-Mining* (2,131b)	2013-10	None
Silkroad2	2013-10-02	23
Vircurex* (1,454b)	2013-10-05	None
GBL	2013-10-26	12
Bips* (1,294b)	2013-11-17	None
Picostocks* (5,896b)	2013-11-29	None
MtGox	2014-02-24	23
Flexcoin	2014-03-02	23
Cryptorush	2014-03-11	23
Mintpal	2014-10-14	23
Silkroad2	2014-11-06	23
Bitstamp	2015-01-04	23, 25
Bter	2015-02-14	23
Cryptsy	2016-01-01	23
Shapeshift	2016-04	23
Gatecoin*	2016-05-13	None
Bitfinex	2016-08-03	12

CRIX

Wrapping up

⊡ A good tool for dealing with unstructured dark data

⊡ Not always possible to compute, so use the best approximations

⊡ Limitations: order of words („bag" of words), order of documents, correlation

What have you seen?

▣ An exploration of a popular cryptocurrency forum

▣ Captured the effect of high profile scamming and hacking

▣ # topics parameter optimal for event detection

▣ An RDC dataset has been created from user posts on bitcointalk.org

▣ The constructed model partitions almost all of the events above a certain severity

CRIX

Dynamic Topic Modeling for Bitcoin Message fora

Marco Linton
Ernie Gin Swee Teo
Elisabeth Bommes
Wolfgang K. Härdle
Cathy Yi-Hsuan Chen

Ladislaus von Bortkiewicz Chair of Statistics

Sim Kee Boon Institute for Financial Economics

International Research Training Group

Humboldt-Universität zu Berlin
lvb.wiwi.hu-berlin.de

www.case.hu-berlin.de

irtg1792.hu-berlin.de

Bayes' rule

⊡ If A and B are events, then Bayes' rule states:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

Six-sided die-tossing example:
- A: an odd number is rolled
- B: a number less than 4 is rolled
- A^c: an even number is rolled

$$P(A|B) = \frac{\frac{2}{3} \cdot \frac{1}{2}}{\frac{2}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2}} = \frac{\frac{1}{3}}{\frac{1}{3} + \frac{1}{6}} = \frac{2}{3}$$

Bayesian machinery

⊡ For point or interval estimation of a parameter in a model M based on data y,

Bayesian inference is based on:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

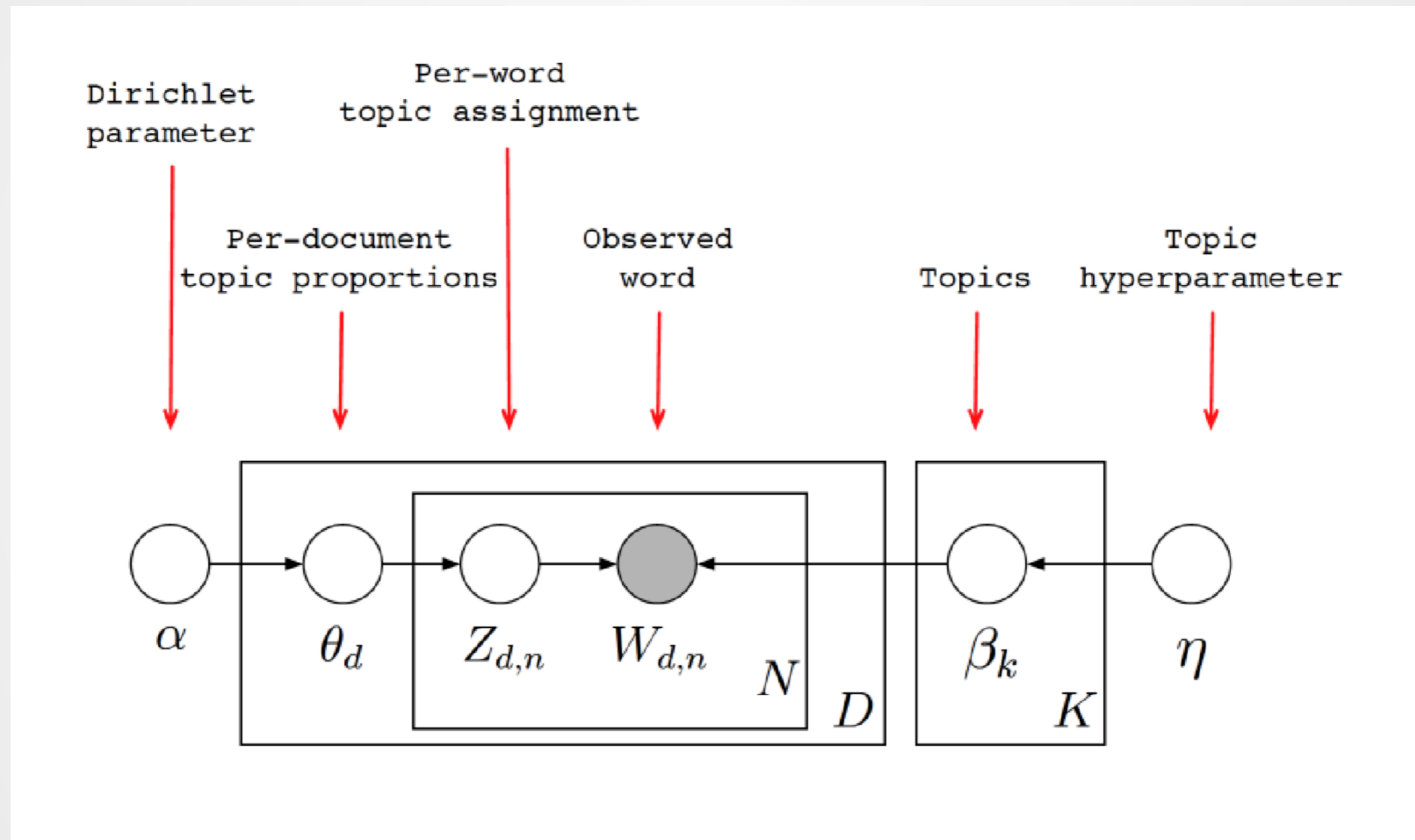$p(\theta)$ is the prior density for the parameter,

$p(\theta|y)$ is the posterior density for the parameter,

$p(y|\theta)$ is the statistical model (or likelihood), and

$p(y)$ is the prior predictive density (or marginal likelihood)

Topic Modelling

▫ Schematic design for the document as a „bag of words"

α and **η** are LDA parameters; for further notation see next slide. Source: D.Blei (2012)

Joint distribution of words and topics

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, \omega_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(\omega_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

$\beta_{1:K}$ topics

$\theta_{1:D}$ per doc topics proportion

$z_{1:D}$ per word topic asignment

$\omega_{1:D}$ observed words

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, \mid \omega_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, \mid \omega_{1:D})}{p(\omega_{1:D})}$$

Bayesian machinery

▣ Disadvantages

- selection of prior
- posterior distributions are heavily influenced by the priors
- high computational cost

▣ hyperparameters need to be checked

▣ requires scenario „robustness tests"

Bayes

Dirichlet distribution

The Dirichlet distribution is defined on a $(k-1)$ dimensional simplex

$$\Delta_k = \left\{ q \in \mathbb{R}^k : \sum_{i=1}^{k} q_i = 1, q_i \geq o, i-1,2,\ldots,k \right\}. \tag{1}$$

It can be thought of as a distribution of random probability mass/density functions (pdf). An excellent example based introduction can be found in Frigyik et al. (2010).

Definition 1 *Let Q be a real value in Δ_k and suppose that $\alpha \in \mathbb{R}^k, \alpha_i > 0$ and defin* $\alpha_0 \overset{def}{=} \alpha^T 1$. *Then Q has a $Dir(\alpha)$ distribution with pdf* $f(q;\alpha) = \dfrac{\Gamma(\alpha_0)}{\prod\limits_{i=1}^{}\Gamma(\alpha_i)} \prod\limits_{i=1}^{k} q_i^{\alpha_i - 1}$.

$$f(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a) + \Gamma(b)} x^{a-1} (1-x)^{b-1}.$$

For $\alpha = (a,b)^T$ with $Q = (X, 1-X) \sim Dir(\alpha)$ for $X \sim Beta(a,b)$.

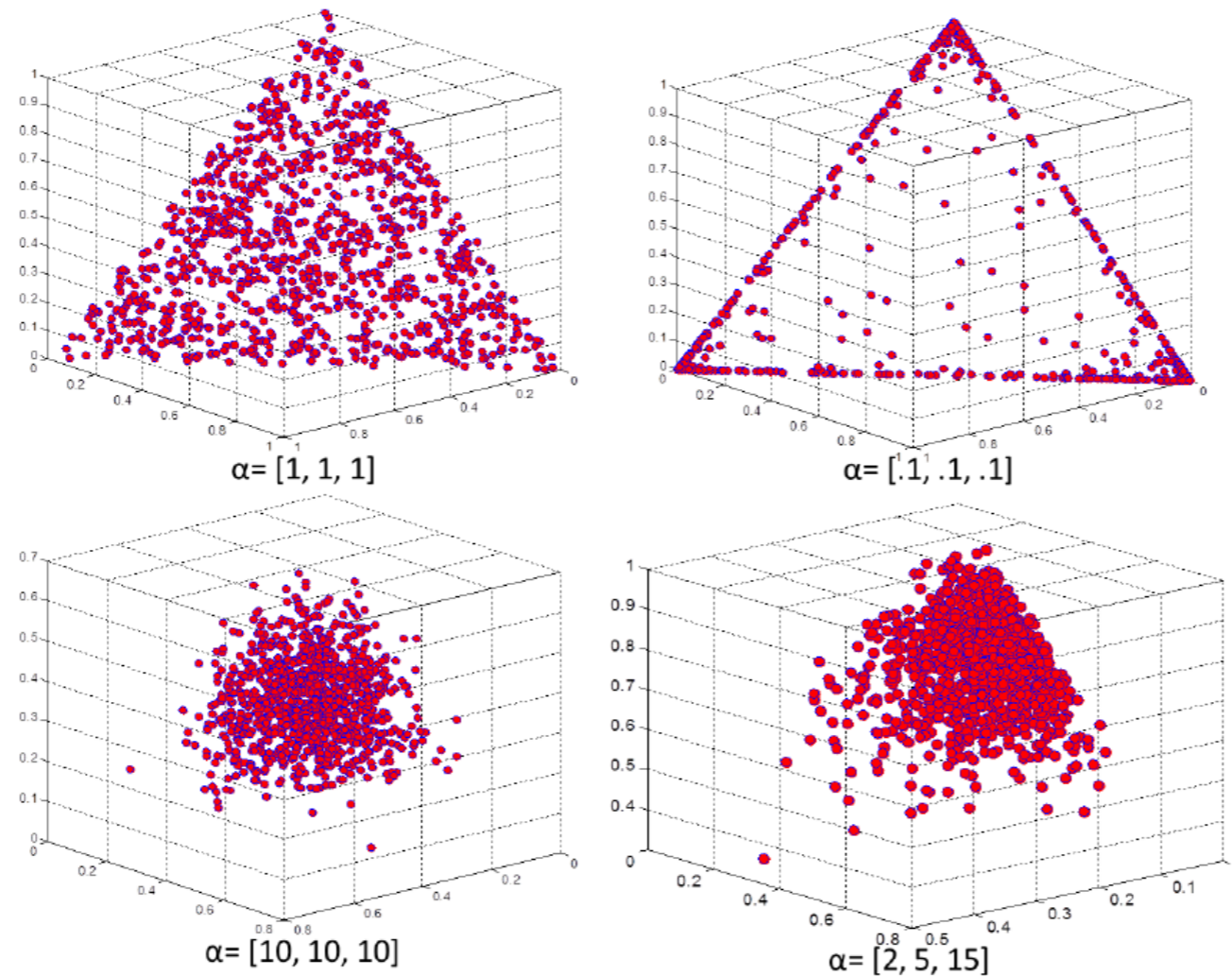# Dirichlet distribution

Figure 1.1: Plots of sample pmfs drawn from Dirichlet distributions for various values of α

XFGdtmDirichlet
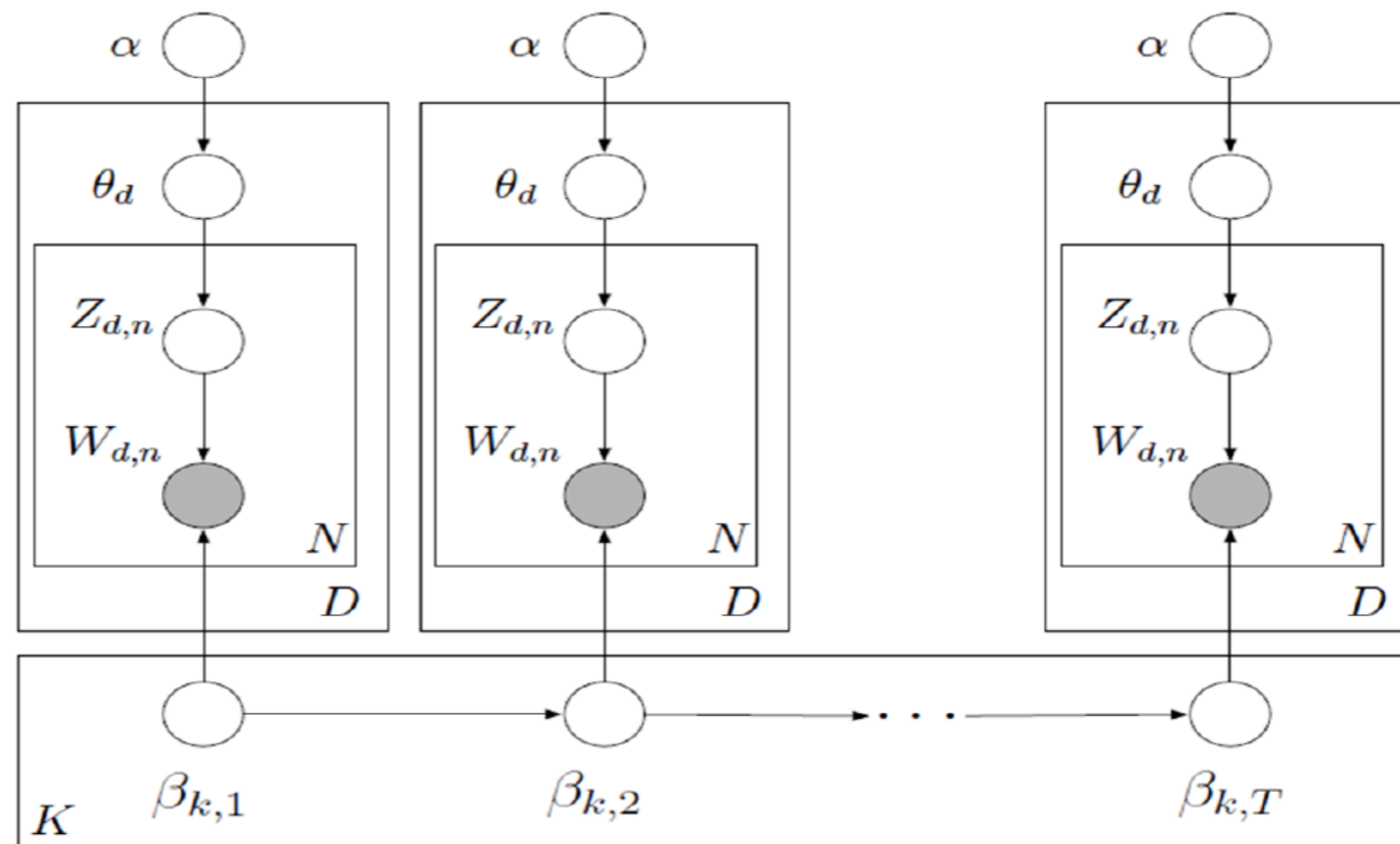
XFGtdmDirichlet

How to do LDA dynamically?

▫ Make the params dynamic

▫ Employ a state space technique

$$\beta_{t,k} | \beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma 2I)$$

$$\alpha_{t,k} | \alpha_{t-1,k} \sim N(\alpha_{t-1,k}, \delta 2I)$$

Figure 3: State space diagram of the dynamic topic model