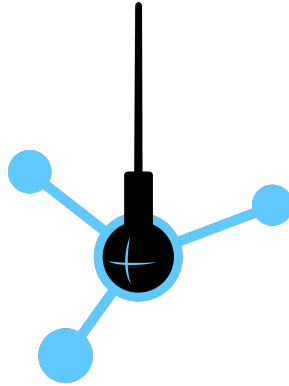


*Final Presentations*

**facebook.tracking.exposed**



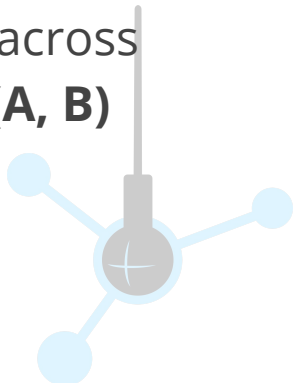
- **User Uniqueness:** How unique is my total post consumption compare to my friends?
- **Post Uniqueness:** Which posts can be considered to have similar content?
- **Investigating Users' Behavior:** What correlation can we find?
- **User Emotional Reaction over time**
- **Cross Bubble Surprise**
- **Content Classification:** can posts be classified under larger topics and how do these topics evolve over time?
- **Exposure and diversity:** how often the algorithm change this? how many people are really appearing in your timeline?

# User Uniqueness

We can use **Jaccard Similarity** to measure uniqueness between any two sets

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

For any two users A and B, we can compare how unique they are across **postIds** or **sources** by using the complement of the similarity **1 - J(A, B)**



# User Uniqueness

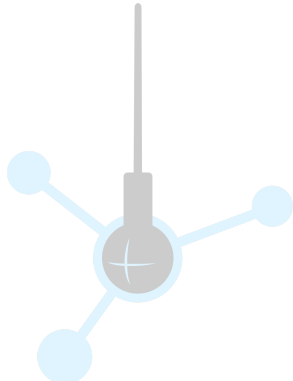
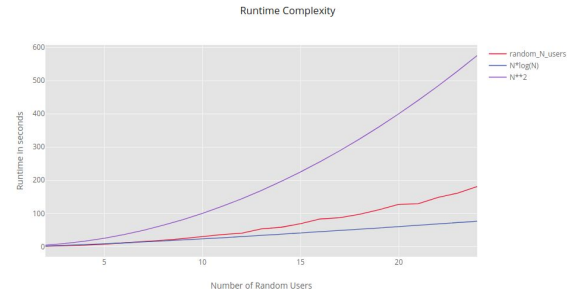
Build a similarity matrix using Jaccard

	chili-basil-tostadas	truffles-persimmon-garbanzo	sushi-dandelion-pickles	lasagne-avocado-onion	parsnip-shawarma-avocado
chili-basil-tostadas	1	0	0	0	0
truffles-persimmon-garbanzo	0	1	0.00211416	0	0
sushi-dandelion-pickles	0	0.00211416	1	0	0.000454064
lasagne-avocado-onion	0	0	0	1	0
parsnip-shawarma-avocado	0	0	0.000454064	0	1

Reduce to a single value for each user

	posts_uniqueness	sources_uniqueness
chili-basil-tostadas	1.000000	1.000000
lasagne-avocado-onion	1.000000	1.000000
sushi-dandelion-pickles	0.999546	0.999546
truffles-persimmon-garbanzo	0.997886	0.997886

Plot the runtime complexity



# Posts Uniqueness

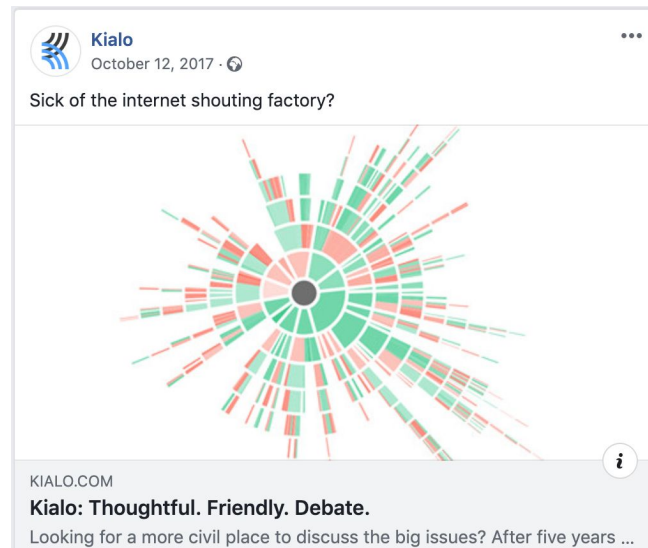
## Data Exploration

The same post can have several source.

Scenario: Kialo posts on a user timeline and the user share it with their friend. The same post will appear as nature=organic and source= user\_name even thou it is a reshare of Kialo'post.

posts various source

cntd_concatText	Source
Sick of the internet shouting factory? Looking for a more civil place to discuss the big issues? After five years of development, we welcome you to Kialo, a system designed for thoughtful debate.	Andrew Garthwaite 1
kialo.com	Antonio Zaccaria 1
	Axel Damelet 1
	Bogdan Vera 1
	Buffalo Nemesi 1
	Claudio Other Claudio 1
	Costas del Sol 1
	Costas Madjammer 1
	Dafne Estesio 1
	Damiano Ramazzotti 1
	Daniele Belleri 1
	Davide Bennato 3
	Elena Casadoro 1
	Enrico Stradaoli 1
	Federica Fulghesu 1
	Flore De Pauw 1
	Giancarlo Sciascia 3
	Guglielmo Papagni 1
	Gustavo Maultasch de Oli.. 1
	Ivy Jelisavac 1
	Jacopo Lanza 2
	Kialo 84

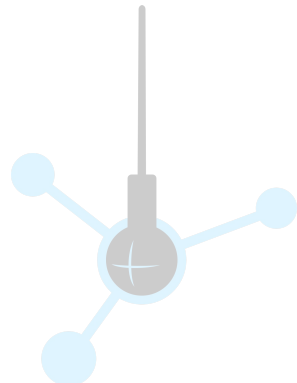


# Posts Uniqueness

## Data Exploration

Concatenated Text	Null	organic	sponsor..
"A strong case could be made for [Gerald] Murnane, who recently turned 79, as the greatest living English-language writer most people have never heard of." <b>Is the Next Nobel Laureate in Literature Tending Bar in a Dusty Australian Town?</b> With the publication of two new books, Gerald Murnane might finally find an American audience.			1
"A strong case could be made for [Gerald] Murnane, who recently turned 79, as the greatest living English-language writer most people have never heard of." With the publication of two new books, Gerald Murnane might finally find an American audience. nytimes.com		1	
"A strong case could be made for [Gerald] Murnane, who recently turned 79, as the greatest living English-language writer most people have never heard of." With the publication of two new books, Gerald Murnane might finally find an American audience. <b>www.</b> nytimes.com			4

Looking at some post from business (The New York Time), the concatenate text can be slightly different, due to AB testing, but the posts are actually the same. We want to consider them as the same content. Even if me and my friend see a different variation of the post, we can consider we consumed the same content.



# Posts Uniqueness

## *Measuring Post Similarity*

```
In [14]: from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import euclidean_distances
vectorizer = TfidfVectorizer()

# This gets a matrix of distances from every text with every other text
# Here lies the memory Problem
def get_similarity_matrix(data):
    X = vectorizer.fit_transform(data.concatenatedText)
    return pd.DataFrame(data=euclidean_distances(X))

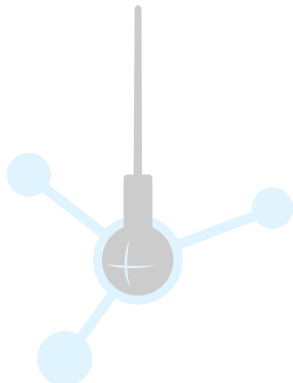
# reduces similarities matrix to the indices of texts below a given threshold
def transform_to_similar_indices(x, threshold):
    return [index for index, value in enumerate(x) if value <= threshold]
```

```
In [ ]: # With this, we can enrich the original data
dist = get_similarity_matrix(df)
## Chosen a threshold of 0.5
df['similars'] = dist.apply(lambda x: transform_to_similar_indices(x, 0.5), axis=1)
# store output:
#df.to_csv('with_similarities3.csv', index=False)
```

**I prepared something to avoid recalculation**

```
In [2]: # reading already enriched dataset:

# need some converters to read that:
from ast import literal_eval
def converter(x):
    return literal_eval(x)
converters={'similars': converter}
df = pd.read_csv('with_similarities3.csv', converters=converters)
```



# Posts Uniqueness

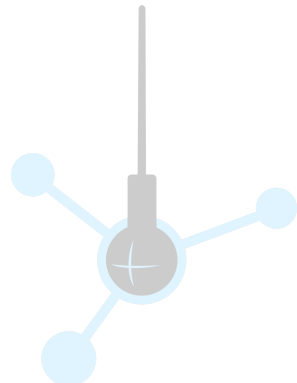
## *Measuring Post Similarity*

```
In [5]: row=3080
all = df.iloc[row]['similars']

df.loc[all,["postId", "concatenatedText"]]
```

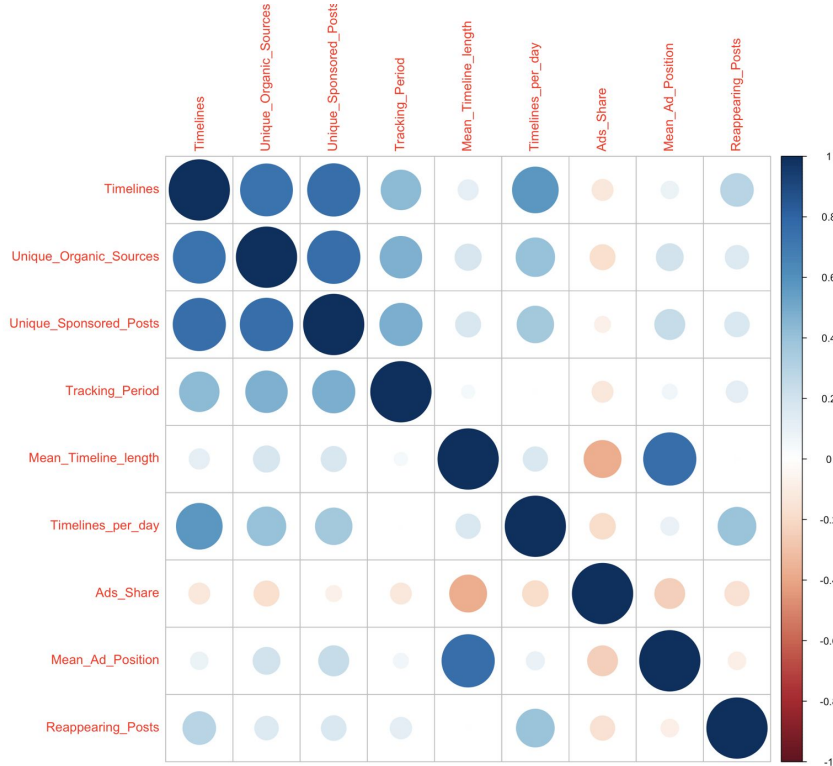
```
Out[5]:
```

	postId	
219	1120121358118429	Sick of the internet shouting factory? Looking for a more civil place to discuss the big issues? After five
1135	1120121358118429	Sick of the internet shouting factory? Looking for a more civil place to discuss the big issues? After five years of d
1161	1120121358118429	Sick of the internet shouting factory? Kialo: Thoughtful. Friendly. Debate. Looking for a more civil place to discuss the big issues? After five
1217	1120121358118429	Sick of the internet shouting factory? Looking for a more civil place to discuss the big issues? After five years of d
1288	1120121358118429	Sick of the internet shouting factory? Kialo: Thoughtful. Friendly. Debate. Looking for a more civil place to discuss the big issues? After five
1912	1120121358118429	Sick of the internet shouting factory? Looking for a more civil place to discuss the big issues? After five years of d
2784	1120121358118429	Sick of the internet shouting factory? Kialo: Thoughtful. Friendly. Debate. Looking for a more civil place to discuss the big issues? After five
3080	1120121358118429	Sick of the internet shouting factory? Kialo: Thoughtful. Friendly. Debate. Looking for a more civil place to discuss the big issues? After five
3136	1120121358118429	Sick of the internet shouting factory? Kialo: Thoughtful. Friendly. Debate. Looking for a more civil place to discuss the big issues? After five
3817	1120121358118429	Sick of the internet shouting factory? Kialo: Thoughtful. F
4120	1139785826151982	Sick of the internet shouting factory? Looking for a more civil place to discuss the big issues? After five years of d
4533	1120121358118429	Sick of the internet shouting factory? Kialo: Thoughtful. Friendly. Debate. Looking for a more civil place to discuss the big issues? After five
5017	1120121358118429	Sick of the internet shouting factory? Looking for a more civil place to discuss the big issues? After five years of d
6469	1120121358118429	Sick of the internet shouting factory? Kialo: Thoughtful. Friendly. Debate. Looking for a more civil place to discuss the big issues? After five
6642	1120121358118429	Sick of the internet shouting factory? Kialo: Thoughtful. Friendly. Debate. Looking for a more civil place to discuss the big issues? After five
6705	1120121358118429	Sick of the internet shouting factory? Kialo: Thoughtful. Friendly. Debate. Looking for a more civil place to discuss the big issues? After five
7105	1120121358118429	Sick of the internet shouting factory? Looking for a more civil place to discuss the big issues? After five years of d
8328	1120121358118429	Sick of the internet shouting factory? Kialo: Thoughtful. Friendly. Debate. Looking for a more civil place to discuss the big issues? After five
8455	1120121358118429	Sick of the internet shouting factory? Kialo: Thoughtful. Friendly. Debate. Looking for a more civil place to discuss the big issues? After five
8496	1120121358118429	Sick of the internet shouting factory? Looking for a more civil place to discuss the big issues? After five years of d
9104	1120121358118429	Sick of the internet shouting factory? Kialo: Thoughtful. Friendly. Debate. Looking for a more civil place to discuss the big issues? After five





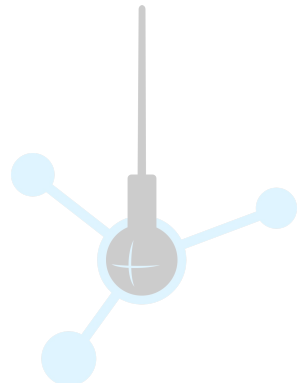
# Investigating Users' Behavior



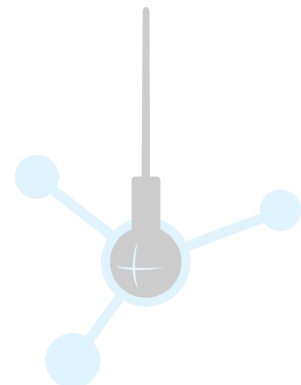
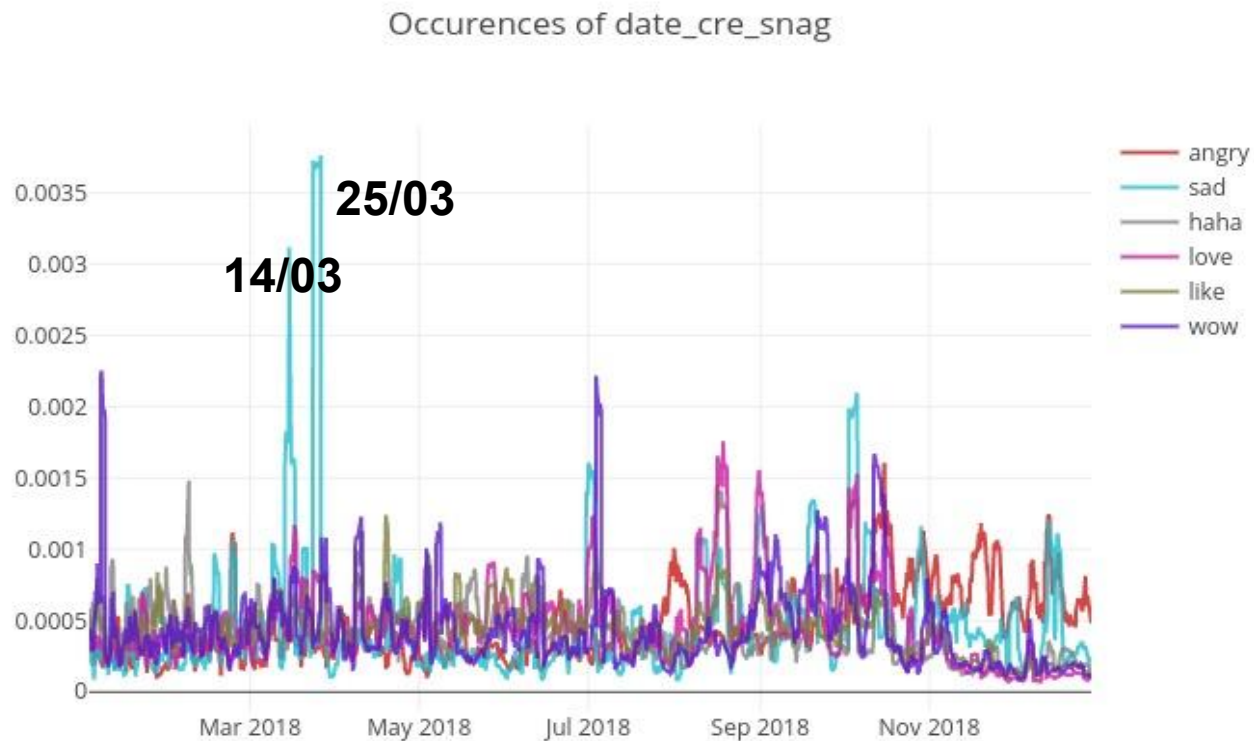
Users are treated similarly by Facebook as long as they have a similar behavior on the platform

At the end, we could cluster users into:

- frequent users (No. of Timelines per Day)
- engaged users (Timeline length)



# Users' Emotional Reactions Over Time



# Example: 14th of March - Real world event

Activities Firefox Web Browser So Apr 28, 11:16

facebook tracking exposed - HTML revision page - Mozilla Firefox

https://facebook.tracking.exposed/revision/ed9541bbe0146bdef7439a8445214463d55dSeed

fbTREX Publications & Talks What we collect and why Know more About Statistics


Summary Metadata Errors Notes

Original HTML snippet

I fucking love science

10 hrs

RIP to one of the greatest minds of our time.



Physicist Stephen Hawking Dies Aged 76

Stephen Hawking, widely regarded as the greatest physicist of our times died peacefully on Wednesday 14th March, his family told the media. He was 76. Hawk

ifscience.com

1.7K Comments

27K Shares

488 Saves

63K43K5.1K113K113K

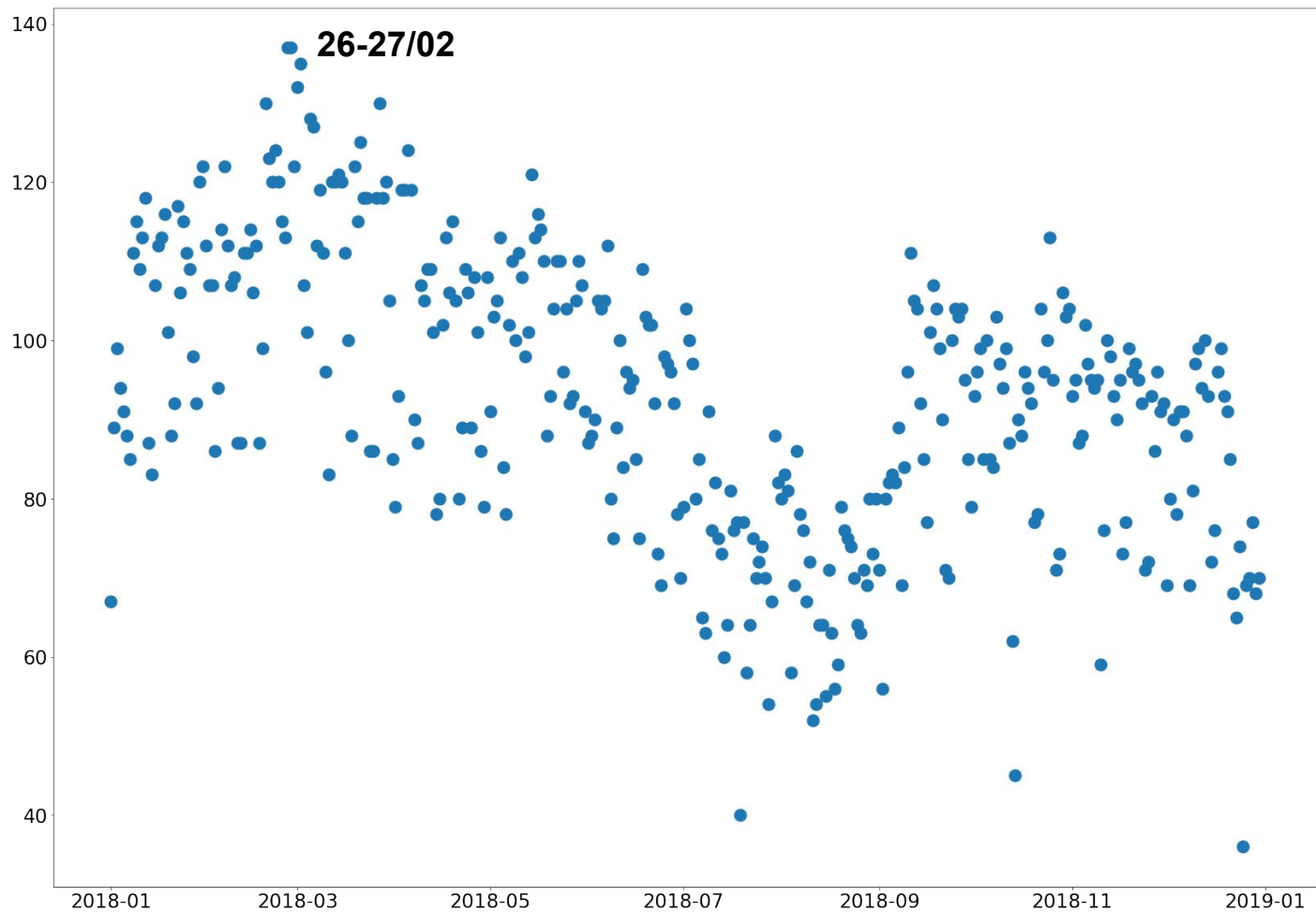
Like

CommentShare

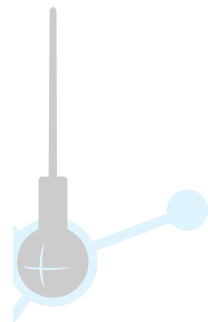
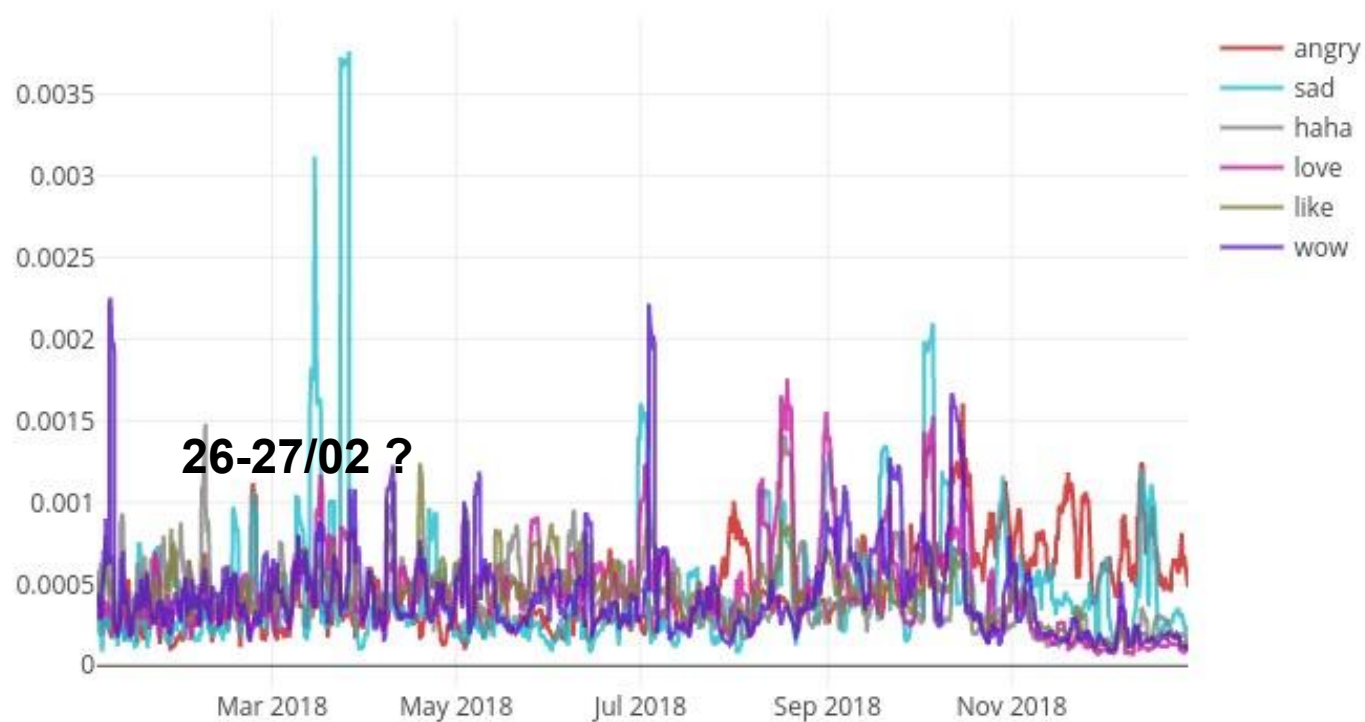
# Example: 14th of March - Random Event

The screenshot shows a Firefox Web Browser window with the title bar indicating the date and time: "So Apr 28, 12:06". The browser's address bar shows the URL: <https://facebook.tracking.exposed/revision/8183d441a0bacb50496b71d2b38710002a22ebbb>. The page content is organized into a header with navigation links: "fbTREX", "Publications & Talks", "What we collects and why", "Know more", "About", and "Statistics". Below the header, there are four main sections: "Summary", "Metadata", "Errors", and "Notes". The "Summary" section is currently active and displays a Facebook post snippet. The snippet text reads: "Vincent Kennes replied to a comment on a post from October 30, 2017." Below this text, there is a "Like Page" button and a "LADbible" link. The snippet also includes a date "October 30, 2017" and the text "'Chewpaca' the alpaca playing chase in the garden". At the bottom of the page, there are links for "19M Views", "LikeShow more reactions", and "CommentShare".

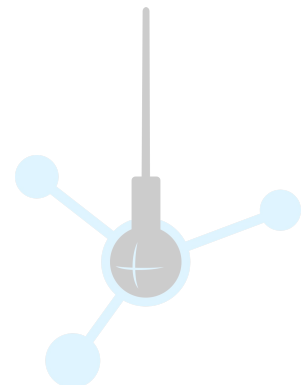
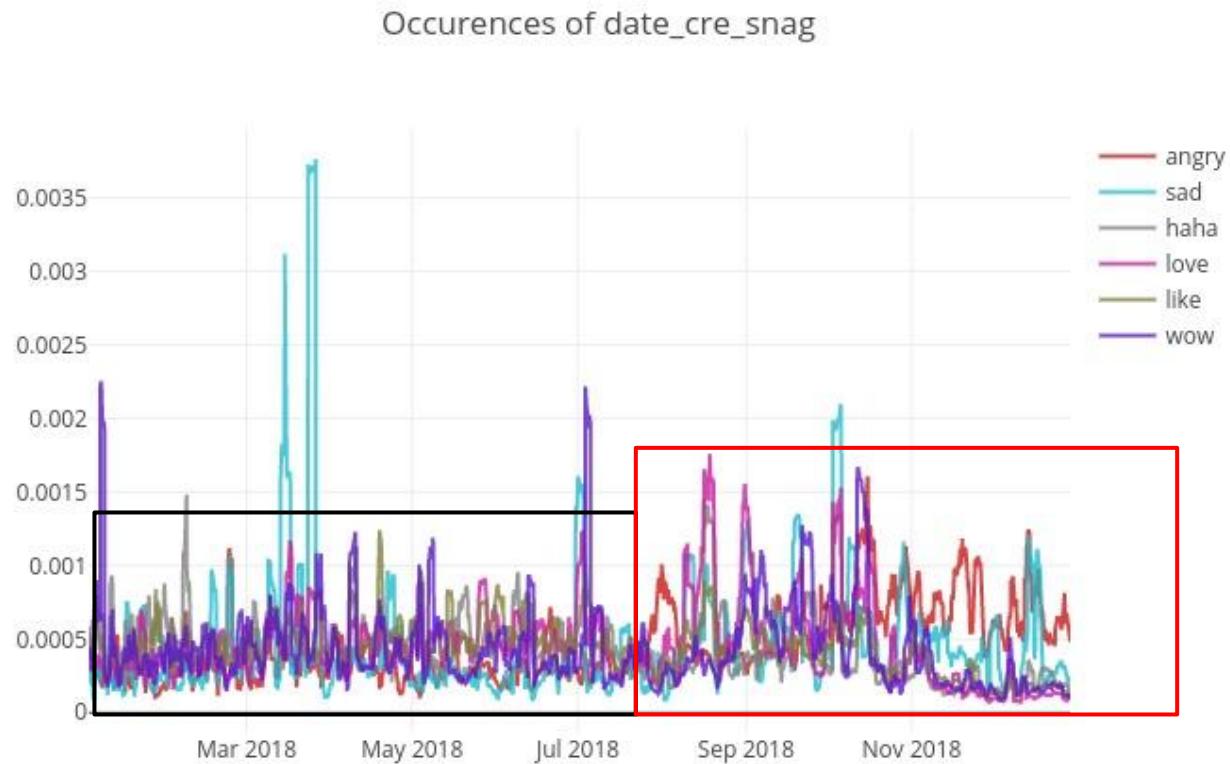
Number of unique logged-in users per day in 2018



## Occurences of date\_cre\_snag

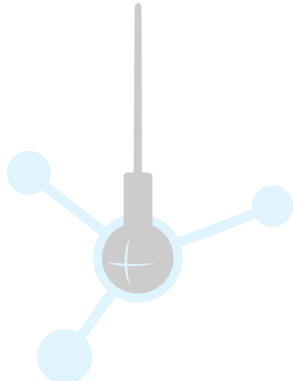


# Anger Levels over time



# Next Steps

- Languages of the posts
- Anger increase in the second part of the year



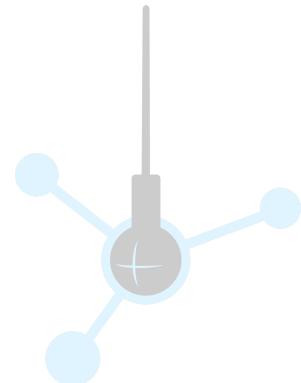
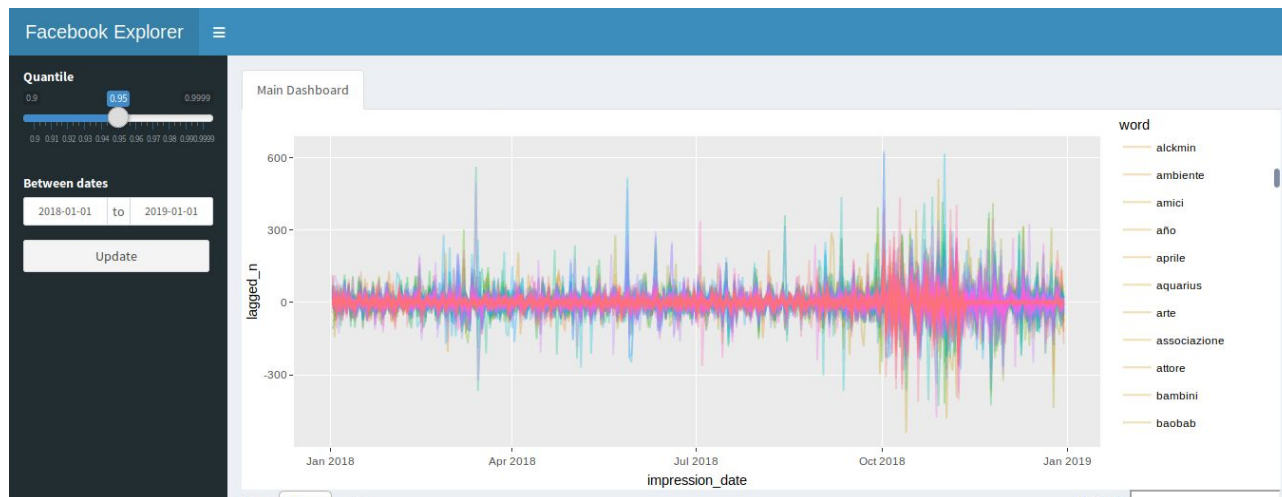


# Cross bubble surprise

## *What news are important for Facebook?*

Main idea is to look at number of impressions containing certain keyword. If a word “spikes” on a certain day, then something happened and we can visualize and analyze it.

To help with visual analysis, I've created a simple tool.

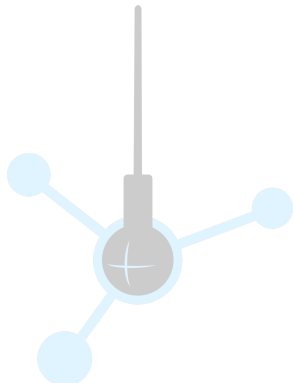


# Cross bubble surprise

## *What news are important for Facebook?*

At the moment, this is only an awareness tool. In future, there are multiple ways how to improve visualization and also analyze actual algorithm:

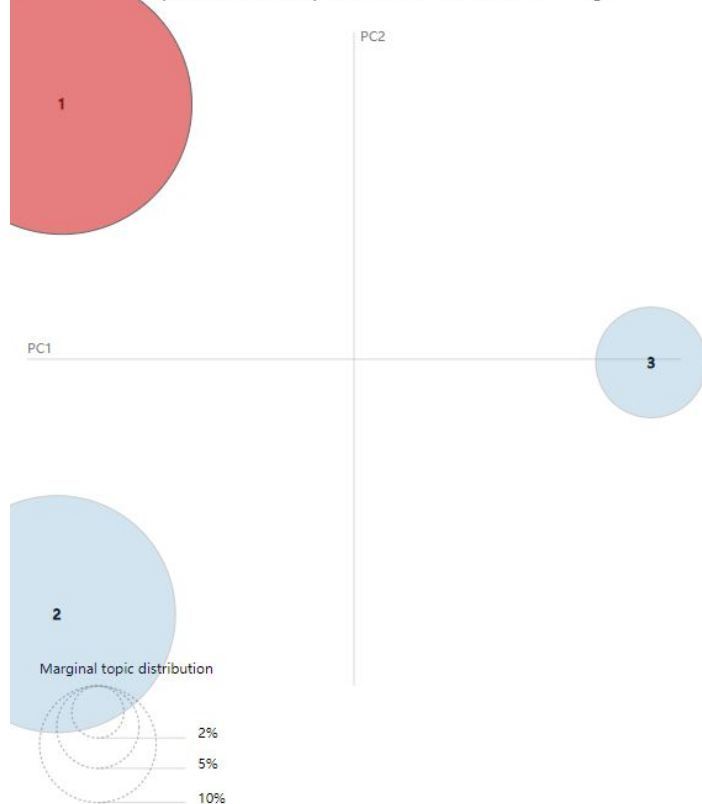
1. Are there different types of news? Are there ways to see organic and manufactured growth?
2. Right now I've looked at 1-day lag. It should be easy to add other summarization functions to help see the patterns better.
3. What news are actually seen by majority of users?
4. Should be interesting to look at ordering of the posts in timelines.
5. Cross-country leaks are not easy to explore since we don't have country of residence for a user. Can be inferred.
6. Compare resulting trends with, e.g., Google Trends to see what news "make it" on Facebook.
  - a. Can also be a way to see what kind of biases are present in data.
7. Summarize by topic (e.g., LDA)/sentiment/etc., not by token.



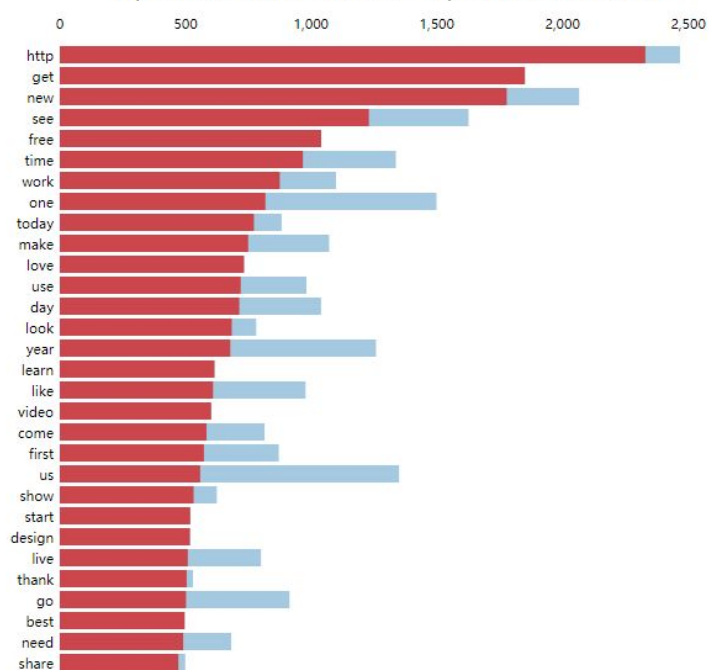
# Content Classification

## Advertisements

Intertopic Distance Map (via multidimensional scaling)



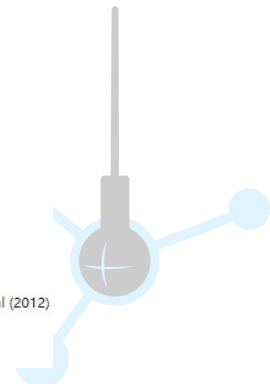
Top-30 Most Relevant Terms for Topic 1 (49.7% of tokens)



Overall term frequency

Estimated term frequency within the selected topic

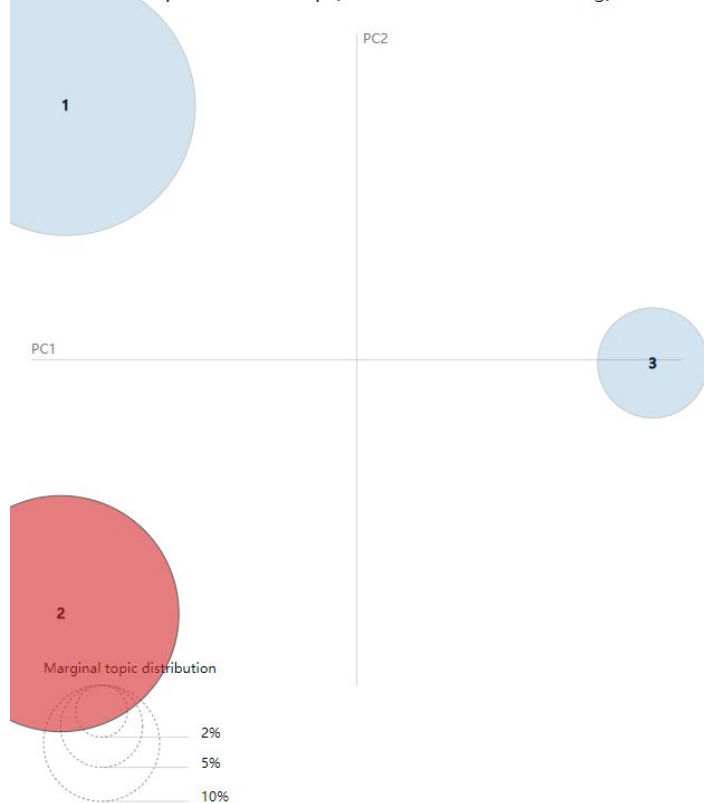
1. saliency(term  $w$ ) = frequency( $w$ ) \*  $\left[ \sum_t p(t | w) * \log(p(t | w) / p(t)) \right]$  for topics  $t$ ; see Chuang et. al (2012)
2. relevance(term  $w$  | topic  $t$ ) =  $\lambda * p(w | t) + (1 - \lambda) * p(w | t) / p(w)$ ; see Sievert & Shirley (2014)



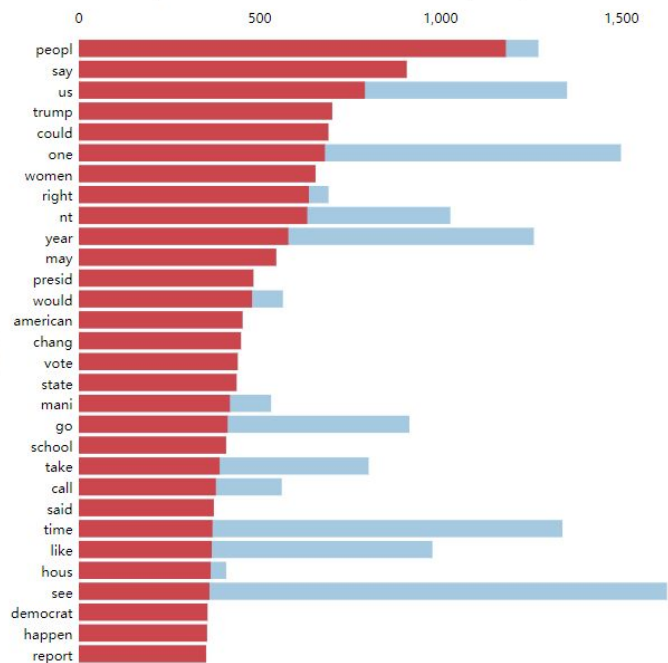
# Content Classification

## Political Posts

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 2 (41.3% of tokens)



Overall term frequency

Estimated term frequency within the selected topic

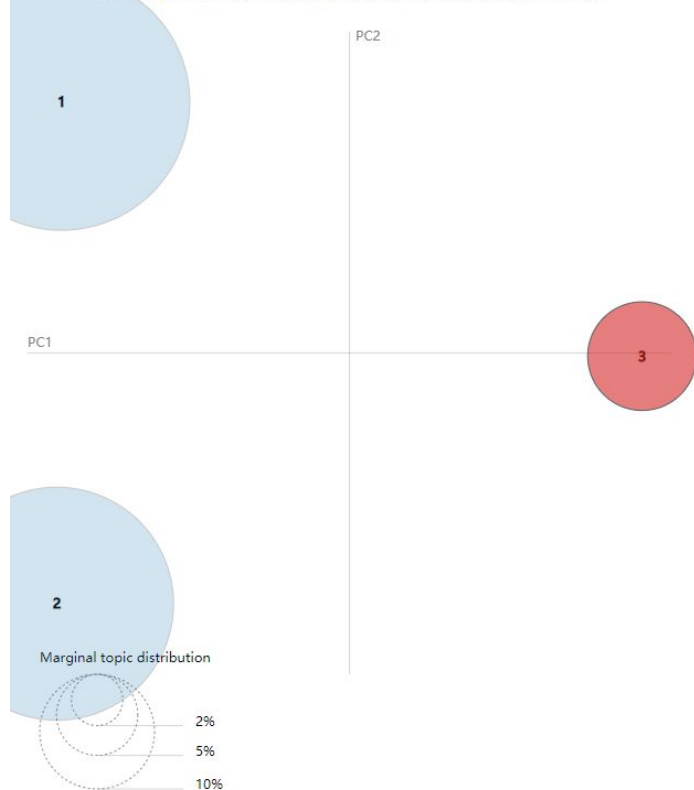
1. saliency(term w) = frequency(w) \* [sum\_t p(t | w) \* log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) =  $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$ ; see Sievert & Shirley (2014)

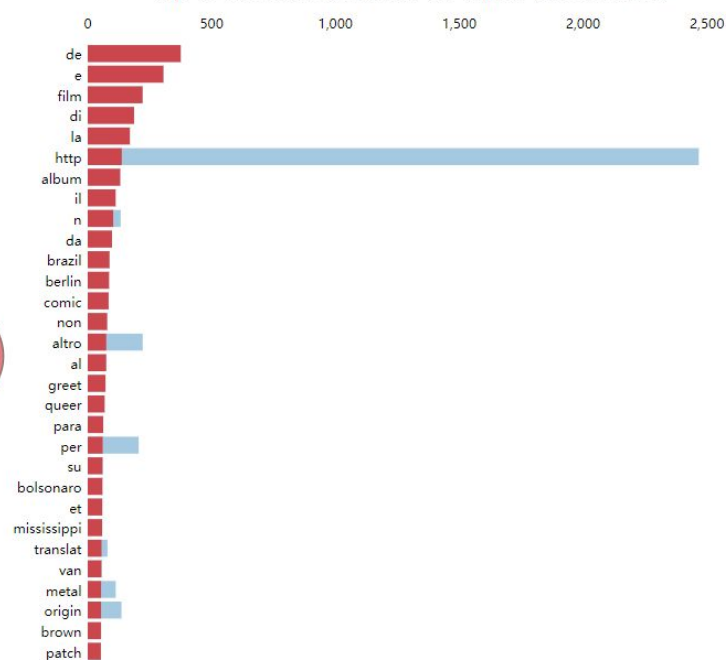
# Content Classification

*Other*

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 3 (9% of tokens)

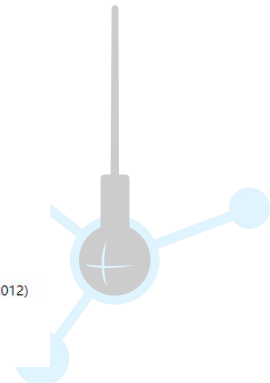


Overall term frequency

Estimated term frequency within the selected topic

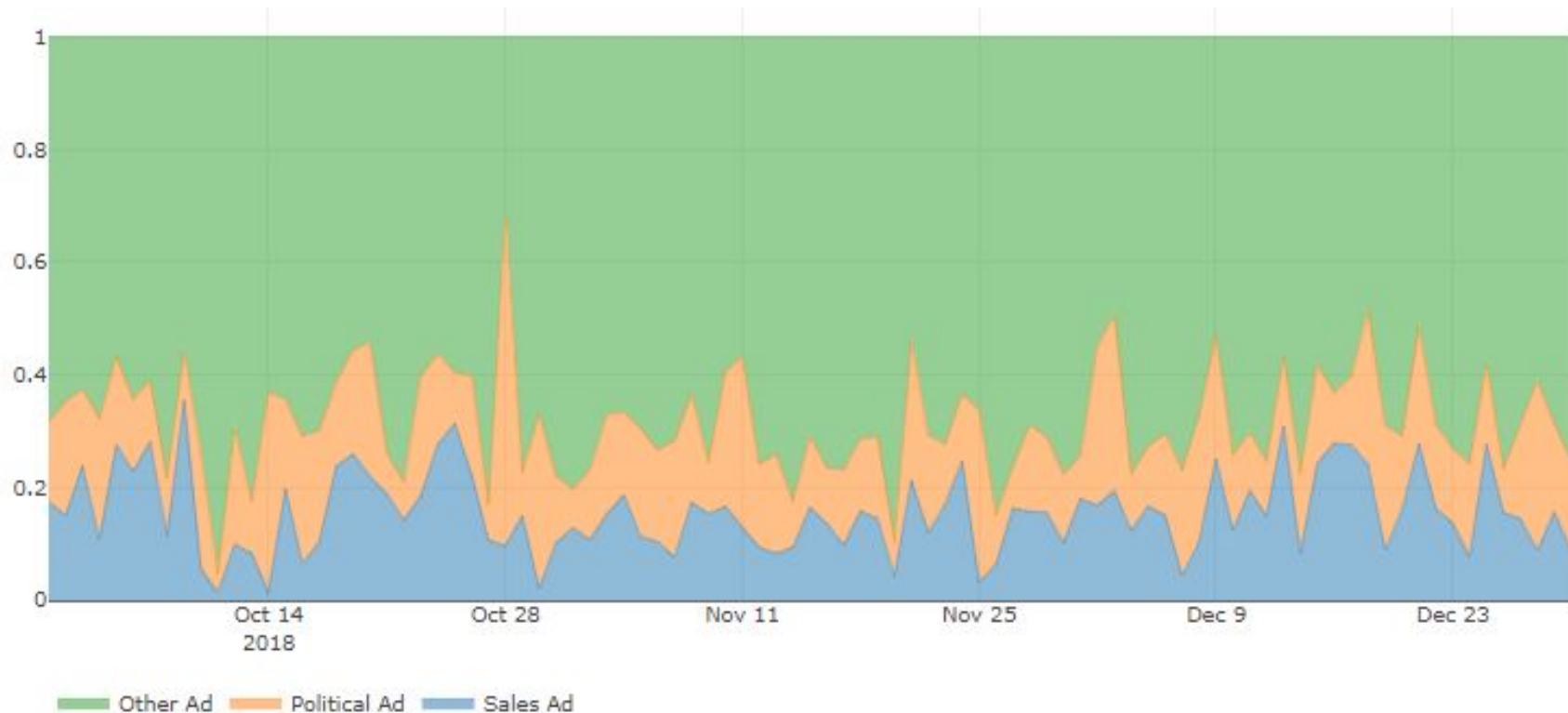
1.  $\text{saliency}(\text{term } w) = \text{frequency}(w) * [\sum_t p(t | w) * \log(p(t | w)/p(t))]$  for topics  $t$ ; see Chuang et. al (2012)

2.  $\text{relevance}(\text{term } w | \text{topic } t) = \lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$ ; see Sievert & Shirley (2014)



# Content Classification

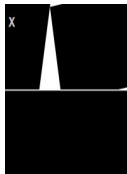
*Sponsored Facebook Post Topics (Oct. - Dec. 2018)*



### *Sponsored Facebook Post Topics (Oct. - Dec. 2018)*

## Political Ad

## Sales Ad



SpSenSsoSredS

35

2 mins

Members of [Spectres](#) are part of a New Year's Eve supergroup tomorrow night at [The Louisiana](#) in Bristol...

The Guardian

GeGspoGnsGertG

66

• 1 Std •

"Trump treats autocrats like friends, and friends like enemies."



Über diese Website  
theguardian.com

In Trump's America, it's important to remember: this isn't normal. | Michael H Fuchs

The breakdown of norms at home undermines democracy, in foreign affairs it undermines security. Americans must hold the president to account

J.Clay

[ScScScScScSc](#)

3 hrs

Great shot by @j\_rago

Socks for your Diadora's [Link in Bio](#)

#jclaysocks #teamjclay #diadora #diadoraofficial #diadoragallery #praisemag  
#minimalmovent #diadoratalk #crecepty #thedropdate #tennissocken  
#strassenmodekultur #mysneakermatch #blkvivs #socken #crecepty #snkrfrkr #hskicks  
#complexkicks #kicksonfire #modernnotoriety #socks #diadoraheritage #outfitsofthe  
#hypebeastkicks #sneakerfreakerofficial #sneakerfreakerfam #instakicks #sneakersnews

# Exposure and diversity

*How many people are populating your newsfeed?*

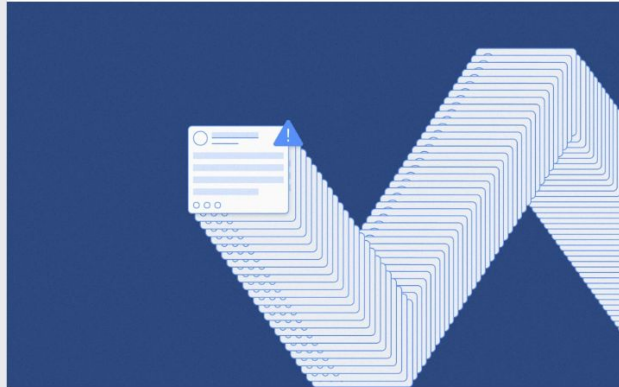
Internet Culture • Analysis

## Facebook isn't restricting your News Feed to 26 friends, no matter what a viral hoax claims

<https://newsroom.fb.com/news/2019/02/inside-feed-facebook-26-friends-algorithm-myth/>

February 6, 2019

**No, Your News Feed Is Not Limited to Posts From 26 Friends**

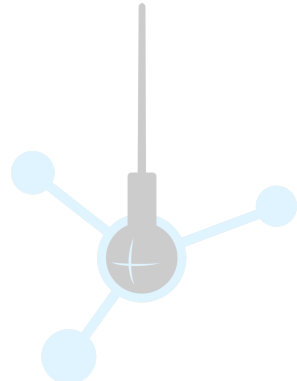


Copy-and-paste memes — those blocks of text posted on message boards, forwarded in emails and shared via social media — are as old as the internet. A recent example started popping up in late 2017 and continues to see the occasional bump in shares. This meme



A version of the most recent meme began spreading on Facebook late last year. (Dado Ruvic/Reuters)

*Let's see if data analysis offer alternative readings*



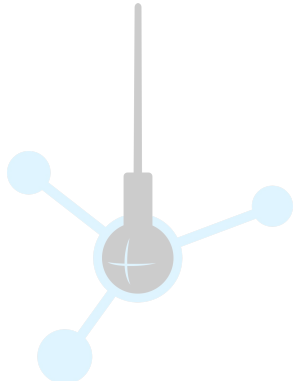


# Exposure and diversity

*How many people are populating your newsfeed?*

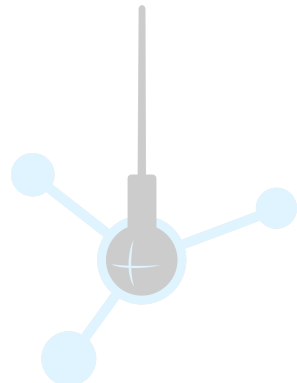
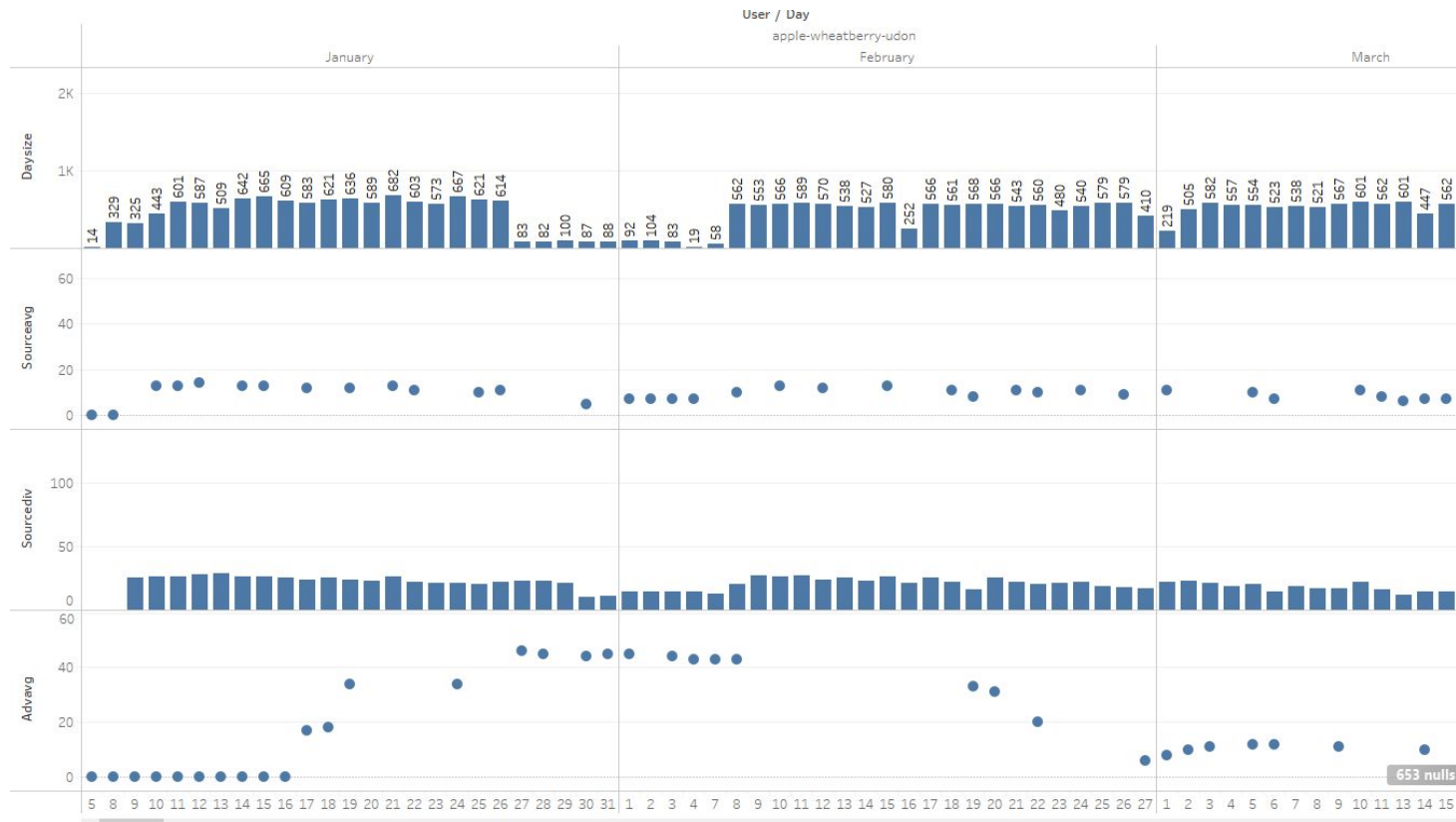
- Considering a queue of 200 posts: how many unique sources appears in your newsfeed?
- Our initial research question was *"can be this an indicator to spot when facebook updates their algorithm?"*

*Let's see if data analysis offer alternative readings*



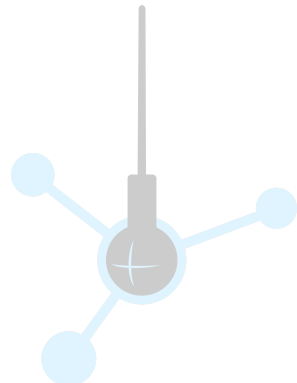
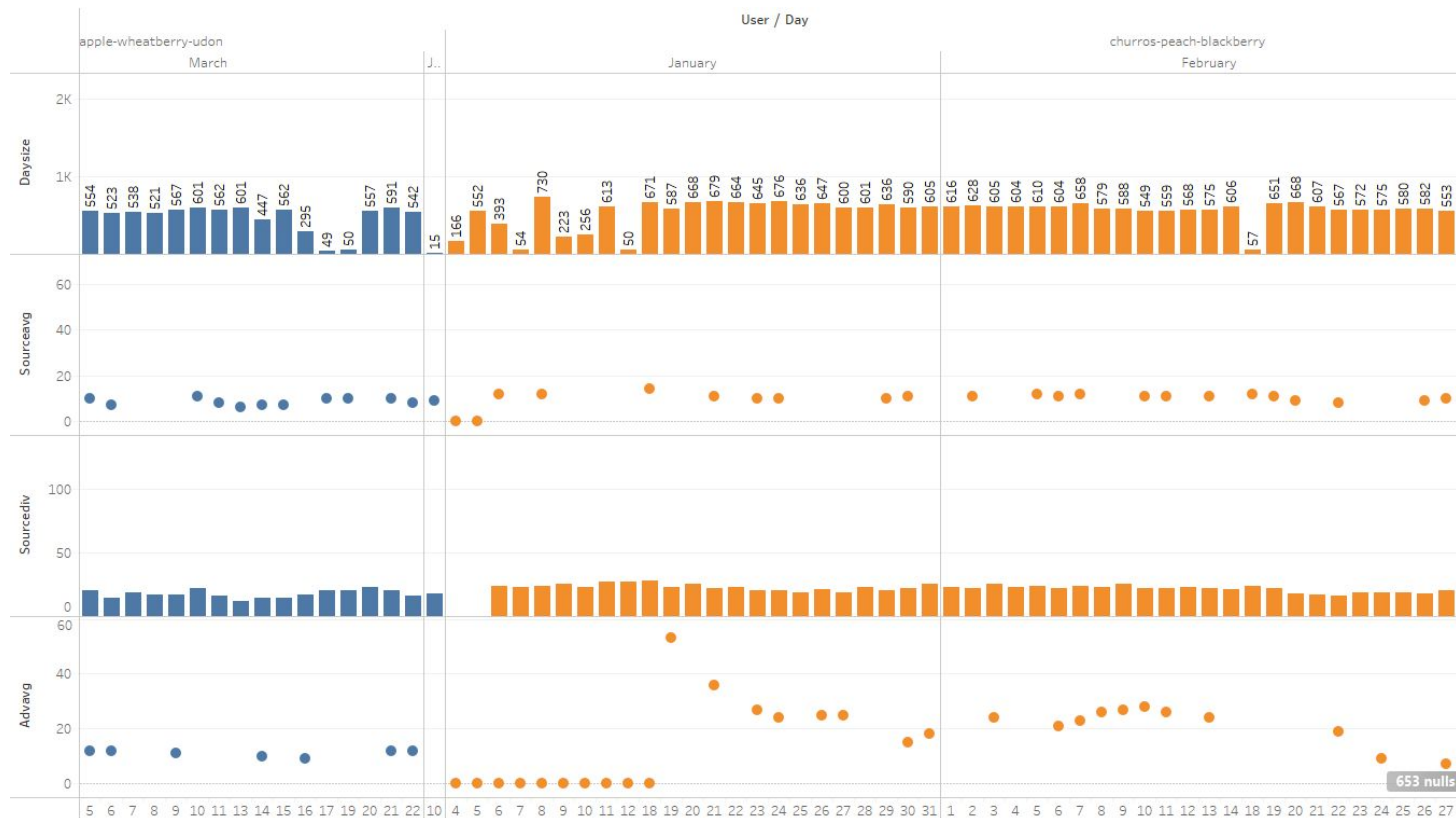
# Exposure and diversity

*How many people are populating your newsfeed?*



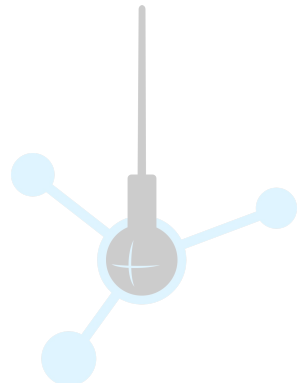
# Exposure and diversity

*How many people are populating your newsfeed?*



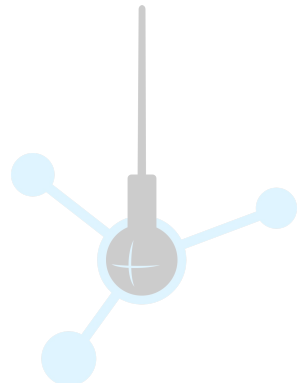
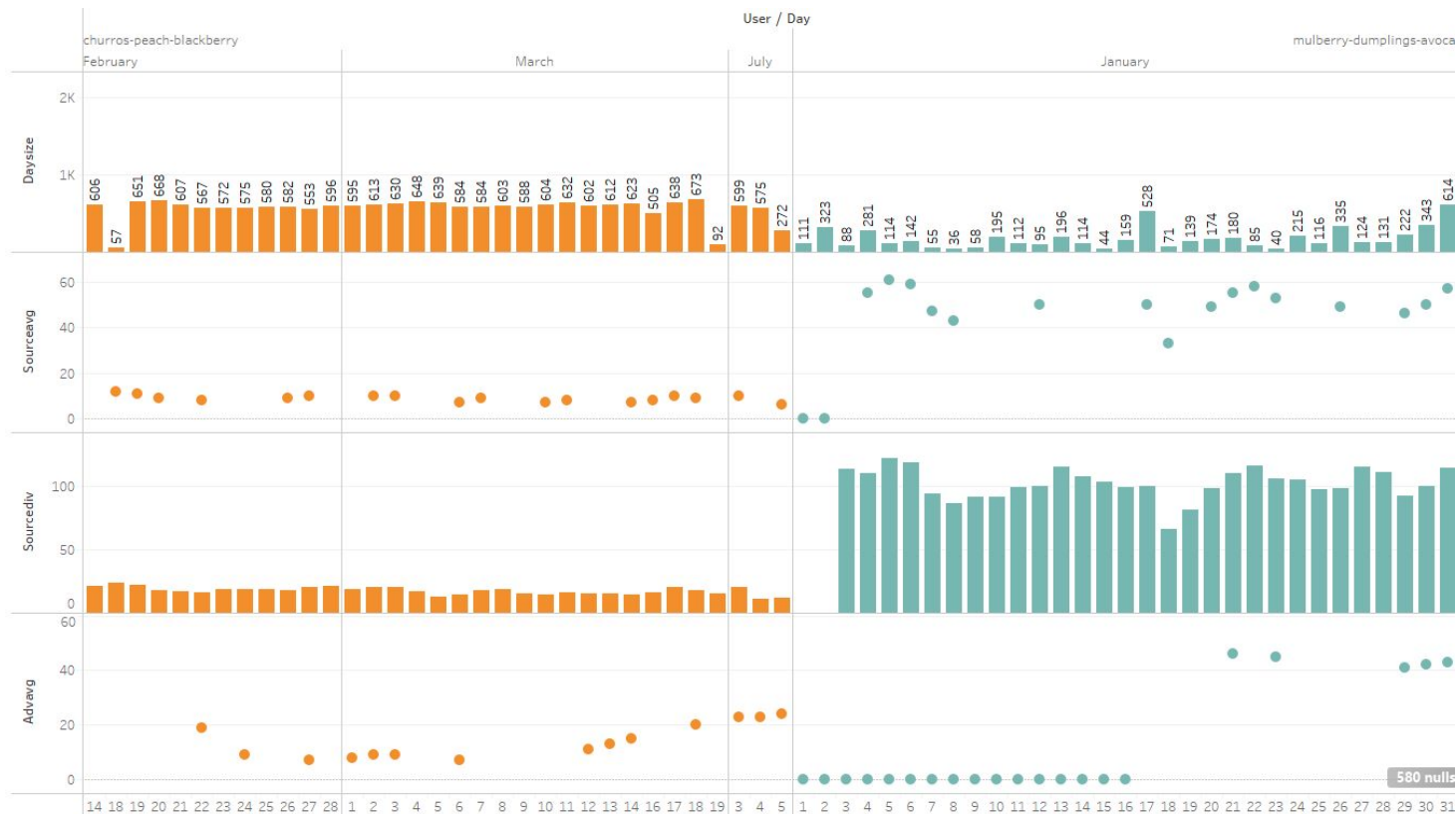
# Exposure and diversity

*How many people are populating your newsfeed?*



# Exposure and diversity

*How many people are populating your newsfeed?*



# Exposure and diversity

*How many people are populating your newsfeed?*

