

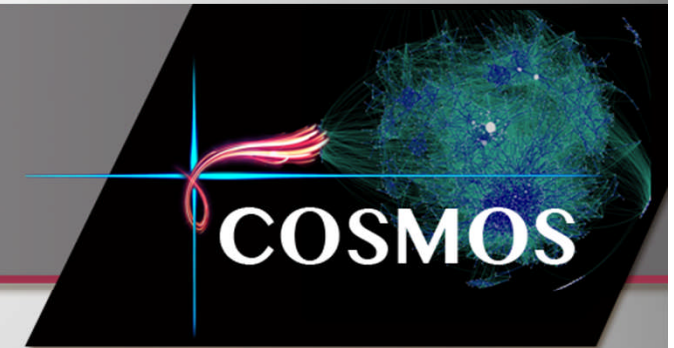
# A Framework for Blog Data Collection: Challenges and Opportunities

Muhammad Nihal Hussain, Adewlae Obadimu, Kiran Kumar Bandeli, Mohammad Nooman, Samer Al-khateeb,  
Nitin Agarwal\*

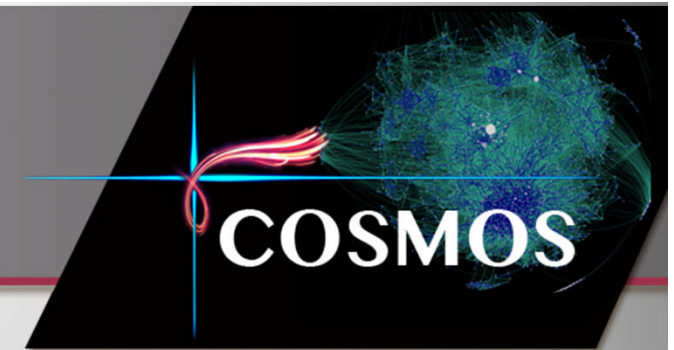
\*Maulden-Entergy Endowed Chair and Distinguished Professor of Information Science  
Collaboratorium for Social Media and Online Behavioral Studies (COSMOS)

[nxagarwal@ualr.edu](mailto:nxagarwal@ualr.edu)

University of Arkansas at Little Rock



- Literature Review
- Data Collection Methodology
- Data Description
- Blogtrackers Analysis
  - Posting Frequency
  - Sentiment Analysis
  - Influence
- Conclusion
- Future Work

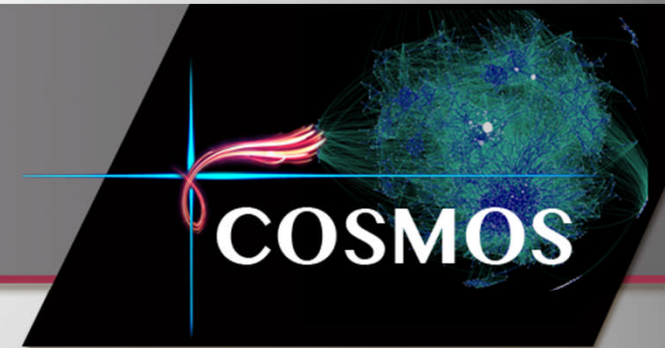


**Blog Crawling:** These crawling techniques scrape only a fraction of the posts from the blogs

- Berger et al. (2011) crawled blogs using RSS feeds and DOM parser.
- Woo et al. (2016) extended the work of Ginsberg et al. (2009) and used blogs to estimate the outbreak of influenza in South Korea.

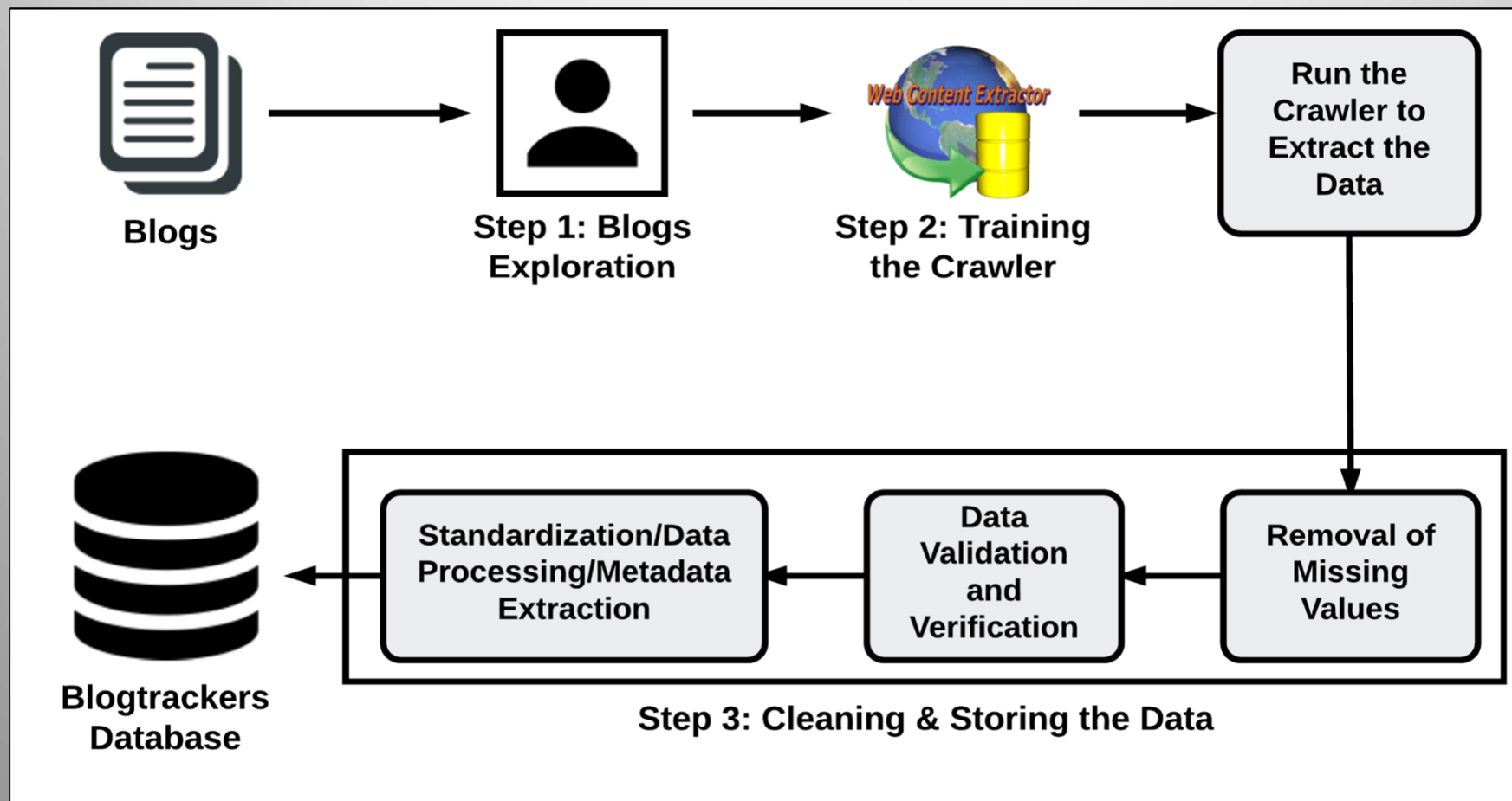
**Deep Web crawling:** These efforts focus primarily on deep web.

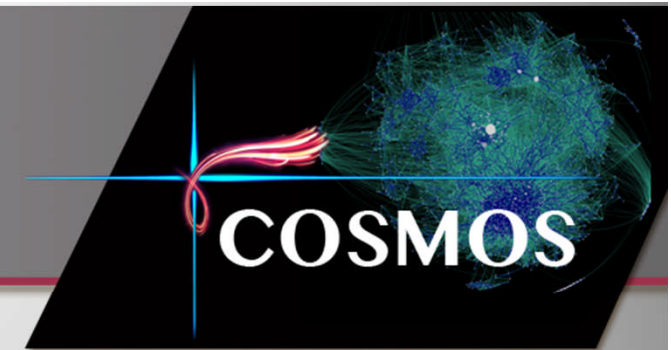
- Zhao et al. (2016) designed a two-stage SmartCrawler framework for scraping the deep web by reverse searching the known deep web sites and ranking the collected sites.
- Wang et al. (2017) developed a crawler that exhaustively scraped textual data from ranked deep web sources with minimal cost.



**Blog tracking tools:** Due to the complexity and inefficiency of blog data collection many blog tracking tools have shutdown.

- BlogPulse – developed by IntelliSeek. provided search and analytical capabilities, automated web discovery for blogs, show the trends of information, and monitor the daily activity. It was shutdown in 2012.
- Blogdex – was developed at “The Media Lab” at MIT. It provided a time-weighted ranking of trending posts. This feature could be used to identify hot-topics in blogosphere. It was shutdown in 2006.
- BlogScope – research product at University of Toronto. Had analytics and Visualization capabilities. Discontinued in early 2012.
- Technorati – used proprietary search and ranking algorithm to provide blog index and search engine. It did not provide blog monitoring or analytical capabilities. Discontinued blog index in 2014.





- Structure of Site
  - Page navigation
  - Archives
- Single-author/Multi-author
- Attributes crawled
  - title
  - author
  - date
  - content
  - comments
  - tags

**NATO Summit in Brussels: Armament and war but also „we will not be silent!“** → title

Posted on May 28, 2017 by line → author

→ date

→ content

At the NATO Summit in Brussels US President Trump wants to collect imaginary debts and the European heads of state reiterate their willingness to spend 2% of GDP on armament. A large wave of armament is coming, for Germany 2% military spending means an increase from 37 billion Euro to 69 billion Euro, for Europe an increase of 200 billion to more than 300 billion Euro. The Europe of crisis and joblessness should pay for preposterous interventions and wars while the military industry rejoices. And there will be cuts in education, science, health care and environment.

The Summit agreed on NATO's participation in the so called "war on IS". In reality this "war on IS" is a series of illegal wars in which Germany now will participate more actively than before. This so called "war on IS" brings terror to a whole region and therefore is provoking terroristic acts. An even more brutal war of bombs will follow with NATO's decision. The majority of the victims are the innocent. Each innocent death strengthens IS and allied "groups of terrorists". The war in and for Syria will be increased turning the whole region into an even more difficult to control tinderbox. The continuous war on terror will also fail because wars do not solve problems but solely increase them by destabilizing whole societies, countries and regions.

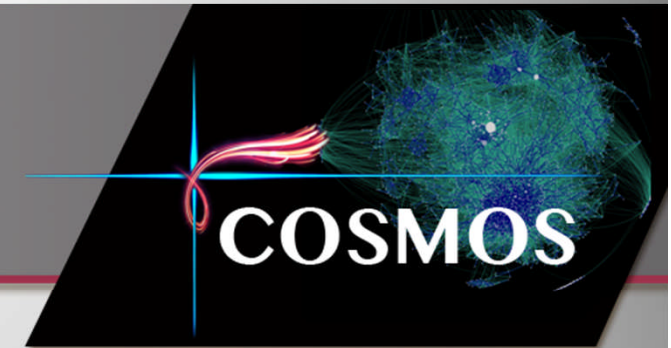
The "3 Cs" of the summit – cash, capabilities, contributions – can be easily translated into more money for modern weapons and for more wars worldwide.

But Brussels was more than a city of the active and cold warriors.

More than 12,000 protesters were marching through Brussels in a colorful, young, broad, impressive and loud demonstration on 24 May. For many hours a trail of peace was moving through the city of Brussels. The actions were creative, musical and international. Many countries with many slogans – an impressive atmosphere. Everywhere "no to NATO" was echoing through the streets. The common demeanor for peace activists, critics of globalization, gender activists, and environmentalists was a clear "no" to further armament. This demand forged a broad coalition of resistance with diverse political backgrounds.

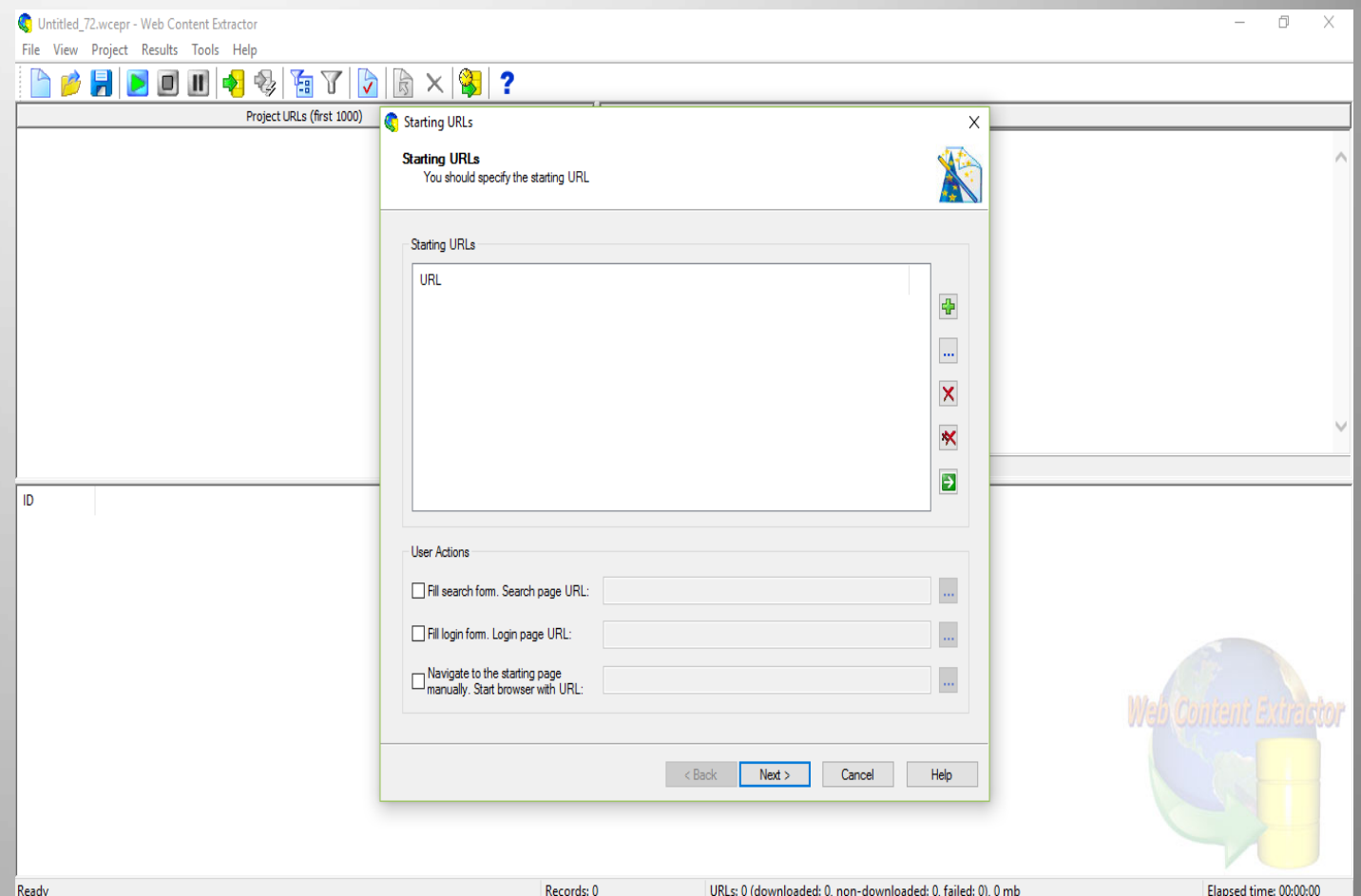
More than 200 people attended to counter summit of the Belgian and international peace movement on 25 May. It was characterized by its internationalism and by discussions on challenges to peace. In an atmosphere of solidarity and mutual tolerance several common points became obvious:

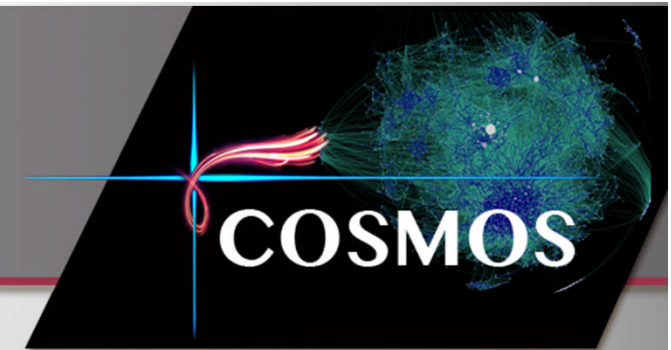
- The peace movements' challenge is to work against the new round of armament towards the 2% of GDP and to struggle for real disarmament: disarmament for development and disarmament for the solution of the social and global challenges. The participants of the conference were prepared to work more intensively towards achieving these challenges.
- The UN ban treaty of nuclear weapons must become reality. There needs to be nuclear disarmament instead of the modernization of nuclear weapons. This is the message to all nuclear weapons states. And Europe finally has to become free of nuclear weapons.
- Cooperation instead of confrontation is not only meaningful but necessary and possible. Especially with Russia. All enemy constructions and bashings serves the preparations of war.
- An end to wars of intervention – in Mali, Afghanistan and many other places of NATO's wars – is the condition and requirement for a peaceful and just development of the world.



To train the crawler:

- Start with seed URLs
- Page navigation techniques
  - Navigation to next page/ next set of posts
  - Navigation to actual/each post
- Extraction pattern
  - DOM pattern for each attribute
- Meta data
  - Permalink
  - Data collection date





## Cleaning using JAVA

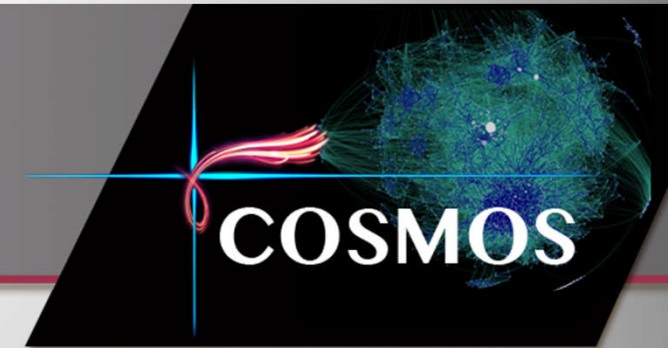
- remove noise from the content
- Data standardization and manipulation

## Data extraction

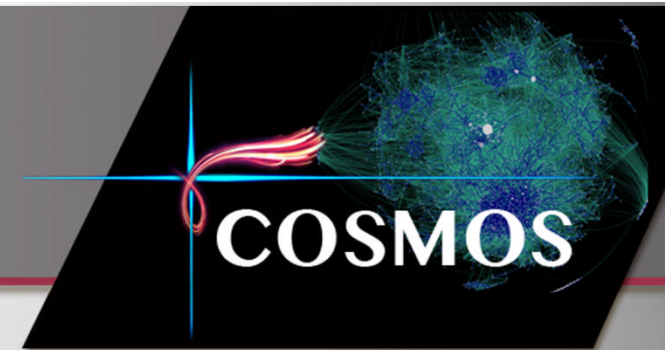
- URLs and Domains using JAVA
- Entity extraction, their types and directed sentiment towards them using AlchemyAPI
- Language of posts using AlchemyAPI
- I.P. address, hosting location, analytics or tracking ID and any related sites using Maltego
- Sentiment and tonality using LIWC 2015







- Challenges faced during data crawling
  - Changing blog structure – Blog site owners can change their blog structure. The crawler need to re-trained for new structure of blog site.
  - Noise – Irrespective of how well a crawler is trained, noise is always crawled. Social media plugins like Facebook share plugins, Twitter share plugins, etc. and advertisements from the blog site could be crawled as JavaScript or HTML code.
  - Limitations of WCE – WCE sometimes fails to crawl dynamic pages that are loaded using JavaScript.
  - Some of technologies used during data collection and extraction steps are not free and require commercial licensing.



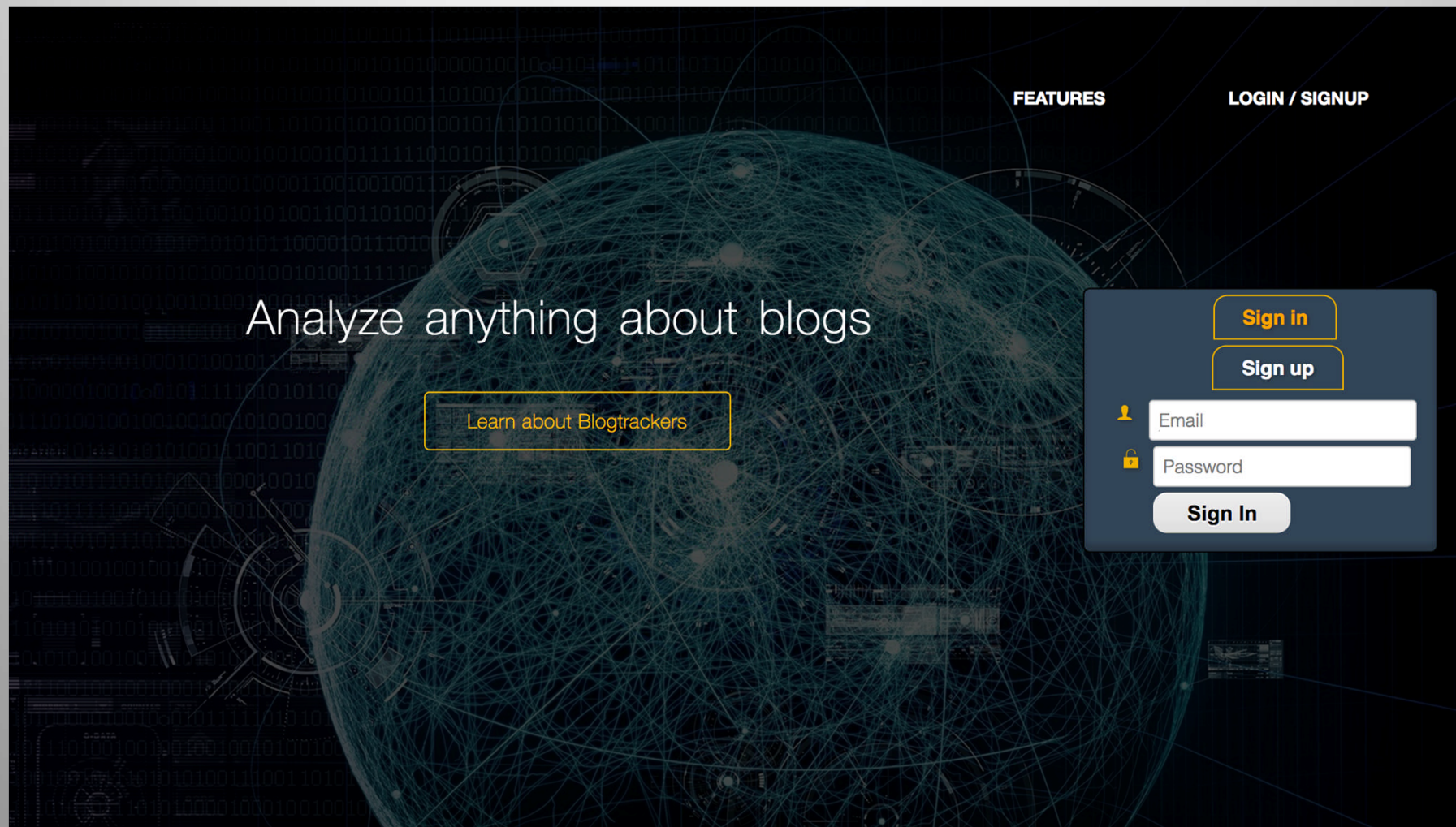
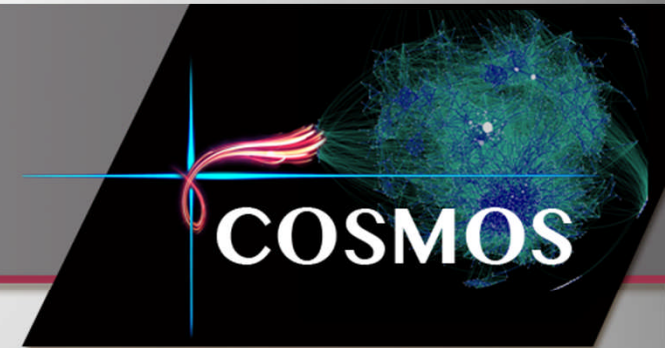
- 257 blogs
- 213,785 blog posts
- More than 1,461,062 links
- 4,058,340 entities
- Earliest post from February 1993
- Last crawled post from May 2017

Language	Blogs	Blog Posts
English	191	164082
Spanish	41	18474
Russian	25	5438
French	25	235
German	22	1411
Italian	16	579
Polish	23	5368
Danish	11	23
Latvian	11	3794
Estonian	7	810

Top 10 languages

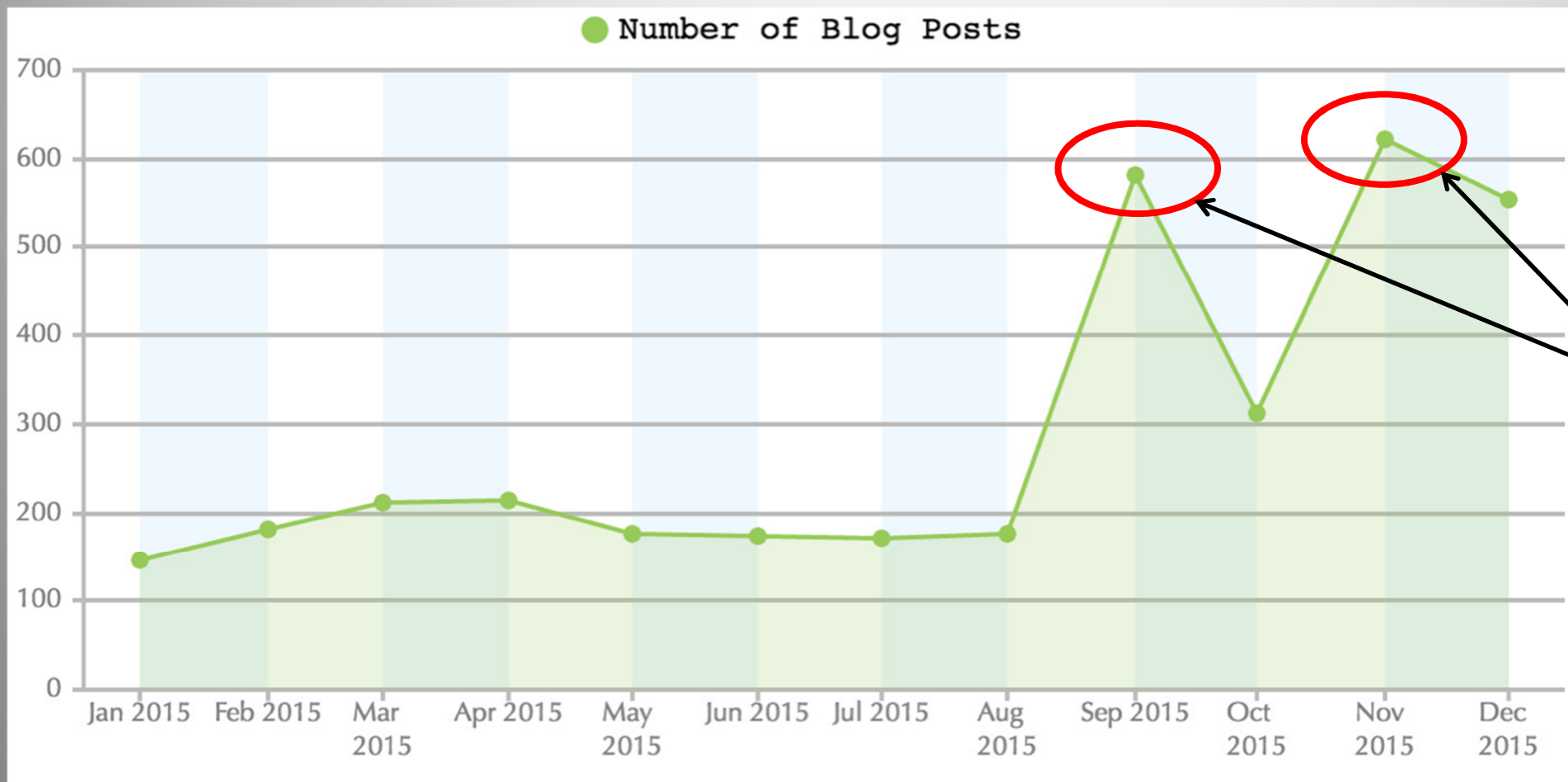
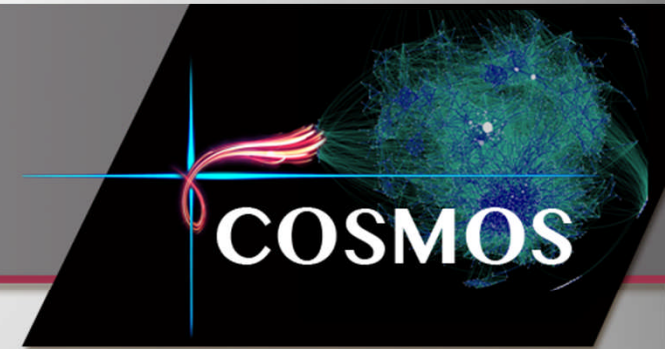
# Blogtrackers

(<http://blogtrackers.host.ualr.edu/>)



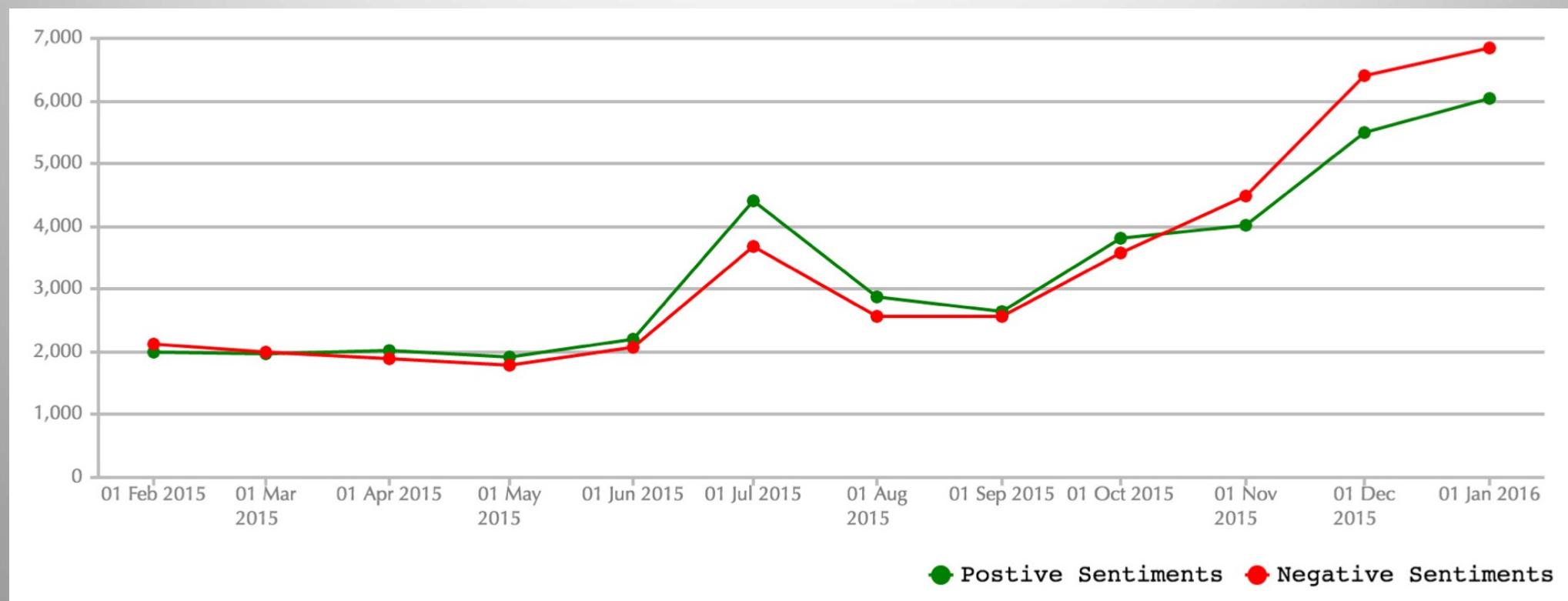
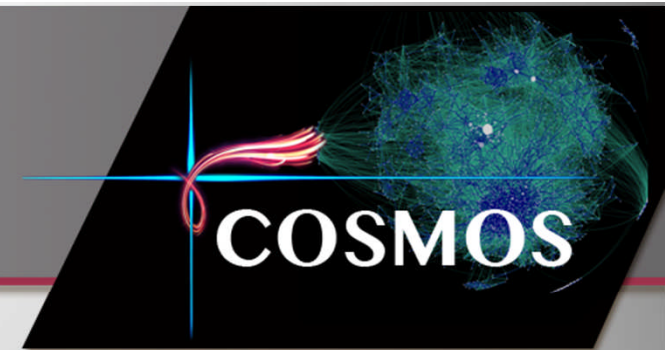
The screenshot shows the main interface of the Blogtrackers website. The background is dark with a complex network visualization of nodes and connections. The text "Analyze anything about blogs" is centered in white. A yellow-bordered button labeled "Learn about Blogtrackers" is positioned below the main text. In the top right corner, there are two links: "FEATURES" and "LOGIN / SIGNUP". A dark grey login/signup form is overlaid on the right side, containing a "Sign in" button, a "Sign up" button, an email input field with a person icon, a password input field with a lock icon, and a "Sign In" button.

# Posting Frequency



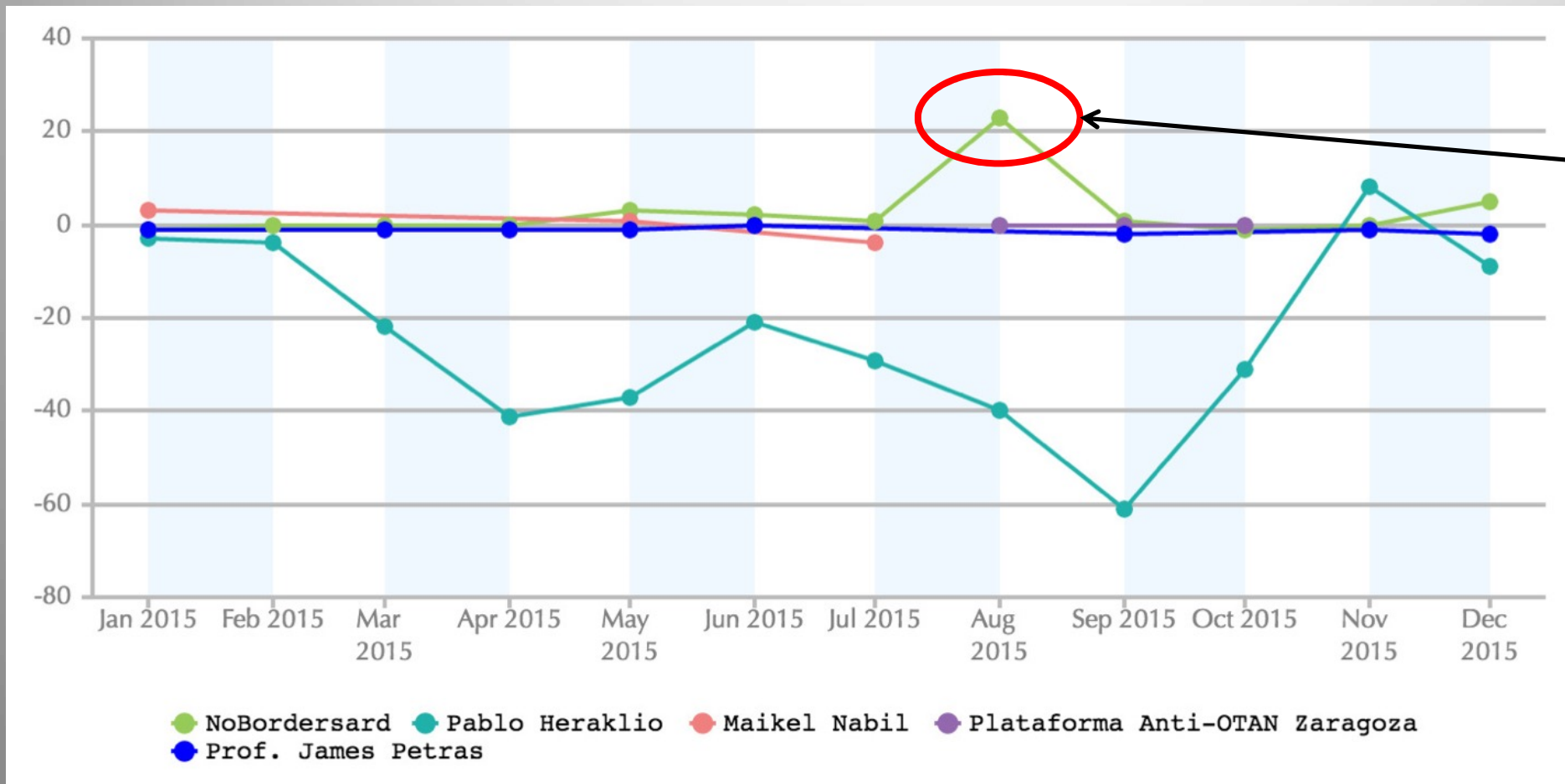
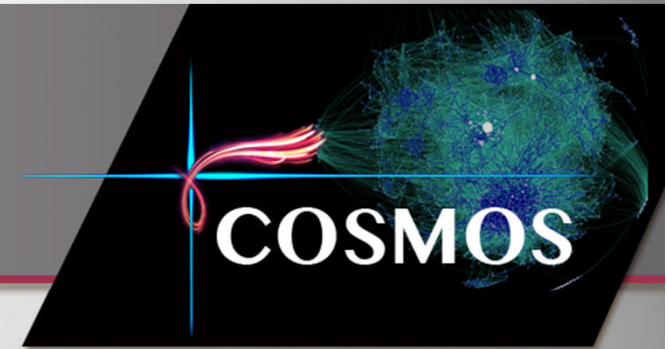
These peaks correspond to intensive blog discussions on NATO's Trident Juncture exercise in late 2015.

Posting Frequency for Anti-NATO blogs from January 2015 to December 2016



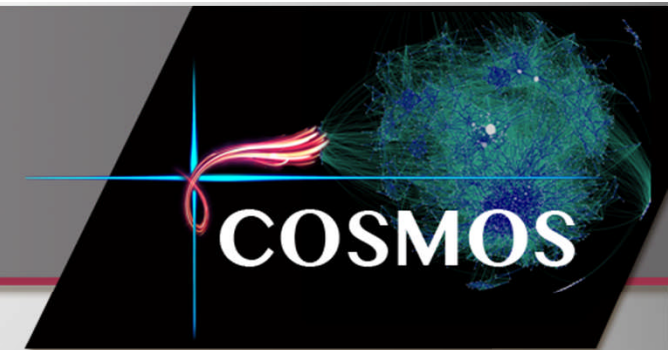
Sentiment trend from January 2015 to December 2015 for Anti-NATO blogs

There is a growing negative sentiment within the blog discourse for NATO's Trident Juncture exercise. The negative sentiment was indicative of the growing anti-NATO narrative in the blogs.



No Bordersard had the highest influence in August 2015

Influence trends for top 5 bloggers



**MOBILITAZIONE CONTRO LA TRIDENT JUNCTURE 2015**



**30 Maggio BENEFIT CASSA ANTIREPRESSIONE SARDA**



**RIGETTATA LA PRIMA RICHIESTA DI SORVEGLIANZA SPECIALE**



**SHO HAI Concerto Anti OTAN**



**AGGIORNAMENTO SULLE SORVEGLIANZE SPECIALI**



**The Battle Between Tradition and Modernity in Egypt**

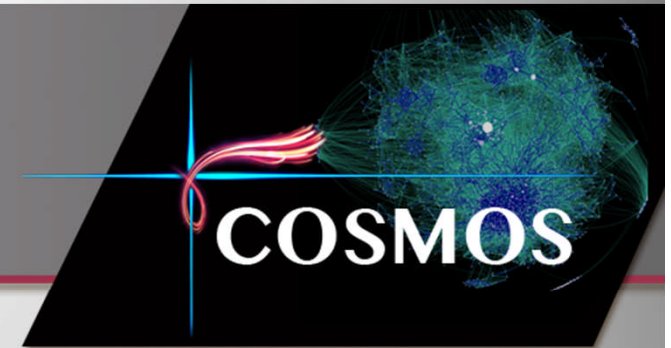


**Indicazioni per arrivare al concentramento della**



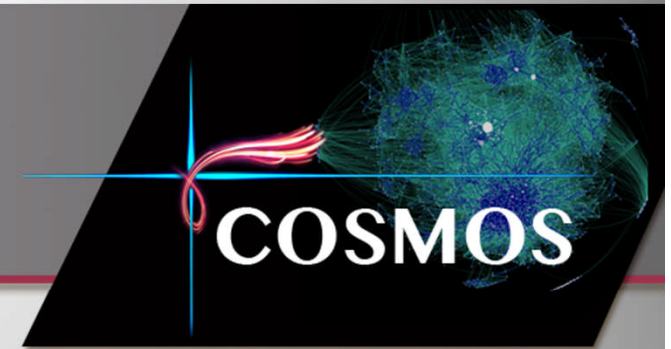
Programma campeggio: OGGI 11 OTTOBRE 2015 MANIFESTAZIONE ANTIMILITARISTA ALLE 18 A CAGLIARI. CONCENTRAMENTO PIAZZA D?ARMI. IL CAMPEGGIO SI TROVA NELLA EX CAVA DI MONTE URPINU (VICINO AGLI ORTI URBANI) VIA RAFFA GARZIA. PER CHI VOLESSE RIMANE L?APPUNTAMENTO STAMANI 9 OTTOBRE, FINO ALLE 11 IN PIAZZA DEL CARMINE. Venerdì 9 ottobre: Dalle 9 alle 11 accoglienza in piazza del carmine ? apertura del campeggio nella ex cava a Monte Urpinu- Nel pomeriggio iniziative in citta? 21.00 cena ? assemblea del campeggio Sabato 10 ottobre 18.00 assemblea sulle prospettive di lotta antimilitarista e contro la trident juncture ? PRESENTAZIONE DEL NUOVO CALENDARIO DELLE ESERCITAZIONI IN SARDEGNA A seguire cena Domenica 11 ottobre Mattina Assemblea conclusiva Pomeriggio corteo Il programma potra? subire variazioni per questioni meteo, per colpa degli sbirri o per imprevisti. PORTA TENDA, SACCO A PELO, PIATTO E POSATE. IL LUOGO DEL CAMPEGGIO VERRA? RESO NOTO DOMANI MATTINA, QUINDI VEDIAMOCI ALL?ACCOGLIENZA!! CAMPEGGIO ANTIMILITARISTA DI LOTTA ? DINTORNI DI CAGLIARI 9-10-11 OTTOBRE 2015 Dalla

Blog posts of No Bordersard (most influential blogger) ranked in decreasing order of influence



- We introduced a semi-Automated crawler, where human trains a crawler tailored for a blog site and lets the crawler run until all the posts from the blog are collected.
- Blogosphere is a relevant information environment to study events, related discourse, and leading narratives.

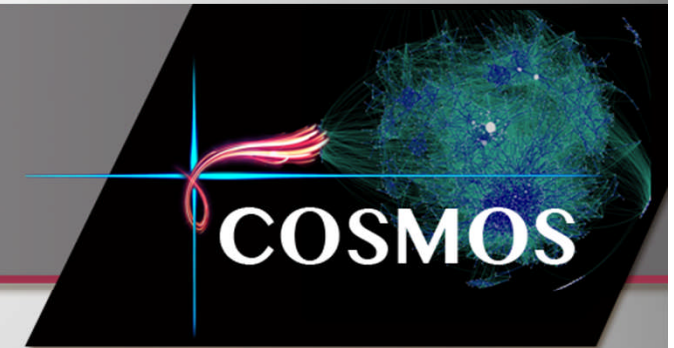




- Automated crawler that identify the attributes from the HTML.
- Add content analysis features to Blogtrackers like:
  - topic modeling,
  - network analysis, and
  - cyber forensics features,
- To not only study blogs individually, but also to understand their coordination structure and information dissemination structure.



# Acknowledgments



This research is funded in part by the

- U.S. National Science Foundation,
- U.S. Office of Naval Research,
- U.S. Air Force Research Lab,
- U.S. Army Research Office,
- U.S. Defense Advanced Research Projects Agency, and
- Jerry L. Maulden/Entergy Fund at the UA-Little Rock.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

Thank you!

[nxagarwal@ualr.edu](mailto:nxagarwal@ualr.edu)

<http://blogtrackers.host.ualr.edu/>