SYMPOZJUM FIZYKI INTERDYSCYPLINARNEJ
W NAUKACH EKONOMICZNYCH I SPOŁECZNYCH

Wątkowa struktura sieci artykułów prasowych

Robert Paluch

Warszawa, 23 czerwca 2016 r.

Dane

The New York Times

Dane od partnerów z **Josef Stefan Institute** – projekt Sophocles.

Ponad **6.5 miliona** artykułów z The New York Times (1851-1999).

Wyłącznie **tytuły i pierwsze akapity** do roku 1987, potem całość.

Zakodowane w postaci **wektorów** [id słowa, liczba wystąpień] (worek słów).

Dane **niekompletne**, wyrwy od kilku dni do kilku miesięcy.

"All the News That's Fit to Print"

The New York Times

LATE CITY EDITION

VOL. CXVIII...No. 40,731 NEW YORK, MONDAY, JULY 21, 1969 10 CENTS

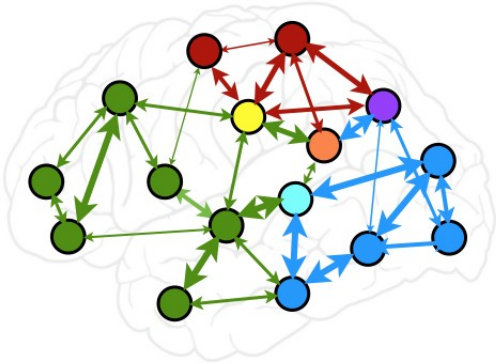
MEN WALK ON MOON

ASTRONAUTS LAND ON PLAIN; COLLECT ROCKS, PLANT FLAG

Voice From Moon: 'Eagle Has Landed'

A Powdery Surface Is Closely Explored

Tworzenie sieci



1. Transformacja wektorów.

a) liczba wystąpień słowa → waga słowa i w dokumencie j

$$(\text{tf-idf})_{i,j} = \text{tf}_{i,j} \times \text{idf}_i$$

$$\text{tf}_{i,j} = \frac{\text{liczba wystąpień słowa } i \text{ w dokumencie } j}{\text{liczba wszystkich słów w dokumencie } j}$$

$$\text{idf}_i = \log_{10} \frac{\text{liczba dokumentów w korpusie}}{\text{liczba dokumentów zawierających słowo } i}$$

← tylko ostatnich **14 dni**

$$\text{idf}_i = 0 \Leftrightarrow \text{idf}_i < 1$$

← Zeruj wagę popularnych słów (występujących w co najmniej **10%** dokumentów)

b) normalizacja wektorów do jedności

Tworzenie sieci

2. Oblicz podobieństwo pomiędzy artykułami (kosinus kąta między wektorami)

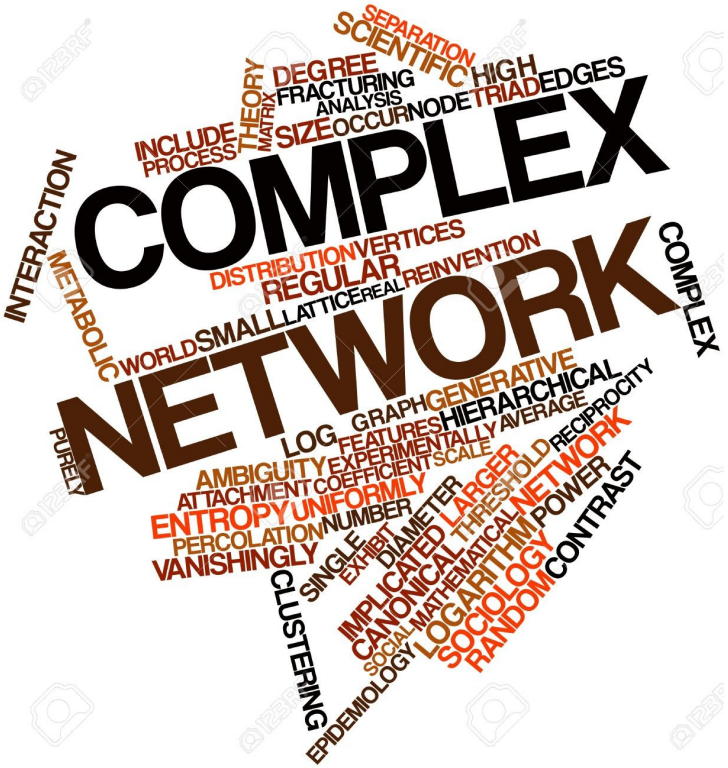
   a) sim(i,j) = 0 $\leftrightarrow$ brak wspólnego słowa o niezerowej wadze,

   b) sim(i,j) = 1 $\leftrightarrow$ identyczne artykuły,

   c) ograniczenie do ostatnich **184 dni**.

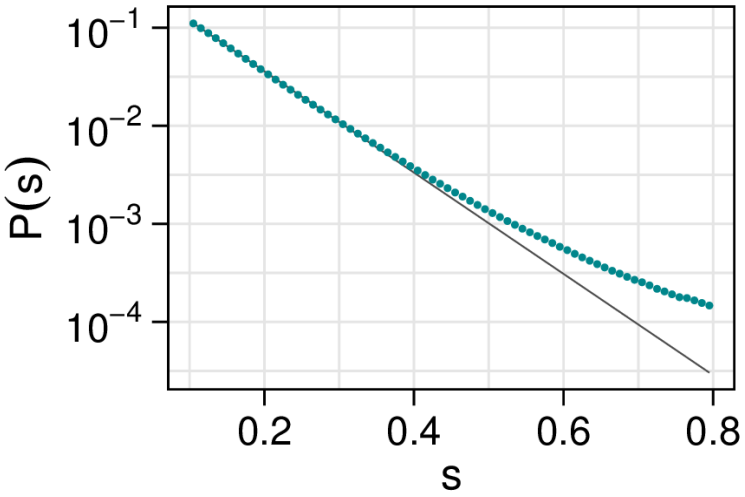
3. Połącz **podobne** artykuły linkiem

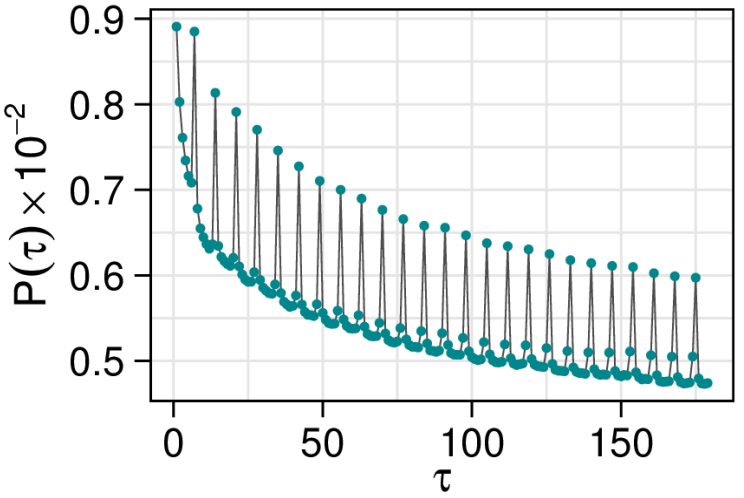
   a) **0.1** ≤ sim(i,j) ≤ **0.8**.



Charakterystyka sieci

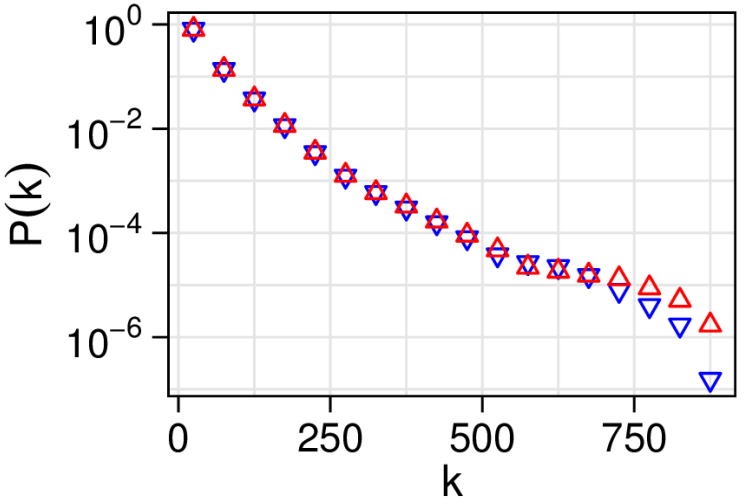
1. Liczba węzłów = **6.45 mln** artykułów
2. Liczba linków = **201.54 mln**
3. Średni stopień = **62.5**
4. Liczba komponentów = **23114**
5. Quasi-wykładniczy rozkład podobieństw
6. Quasi-potęgowy rozkład odstępów czas.
7. Szerokie rozkłady stopni wierzchołków
8. Szeroki rozkład wielkości komponentów

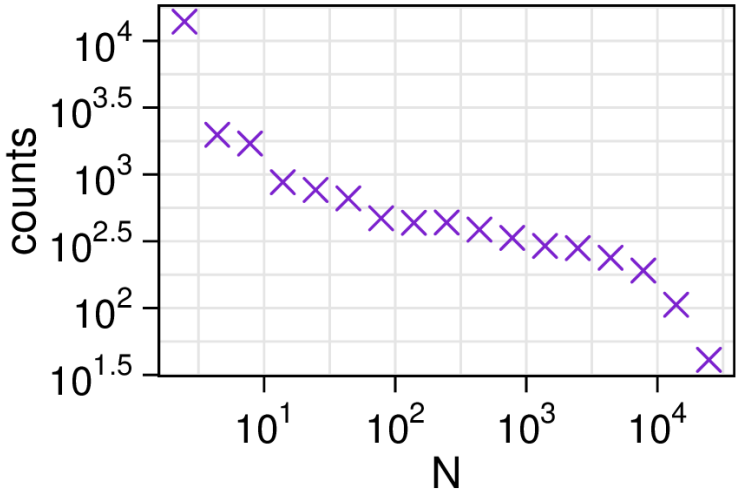




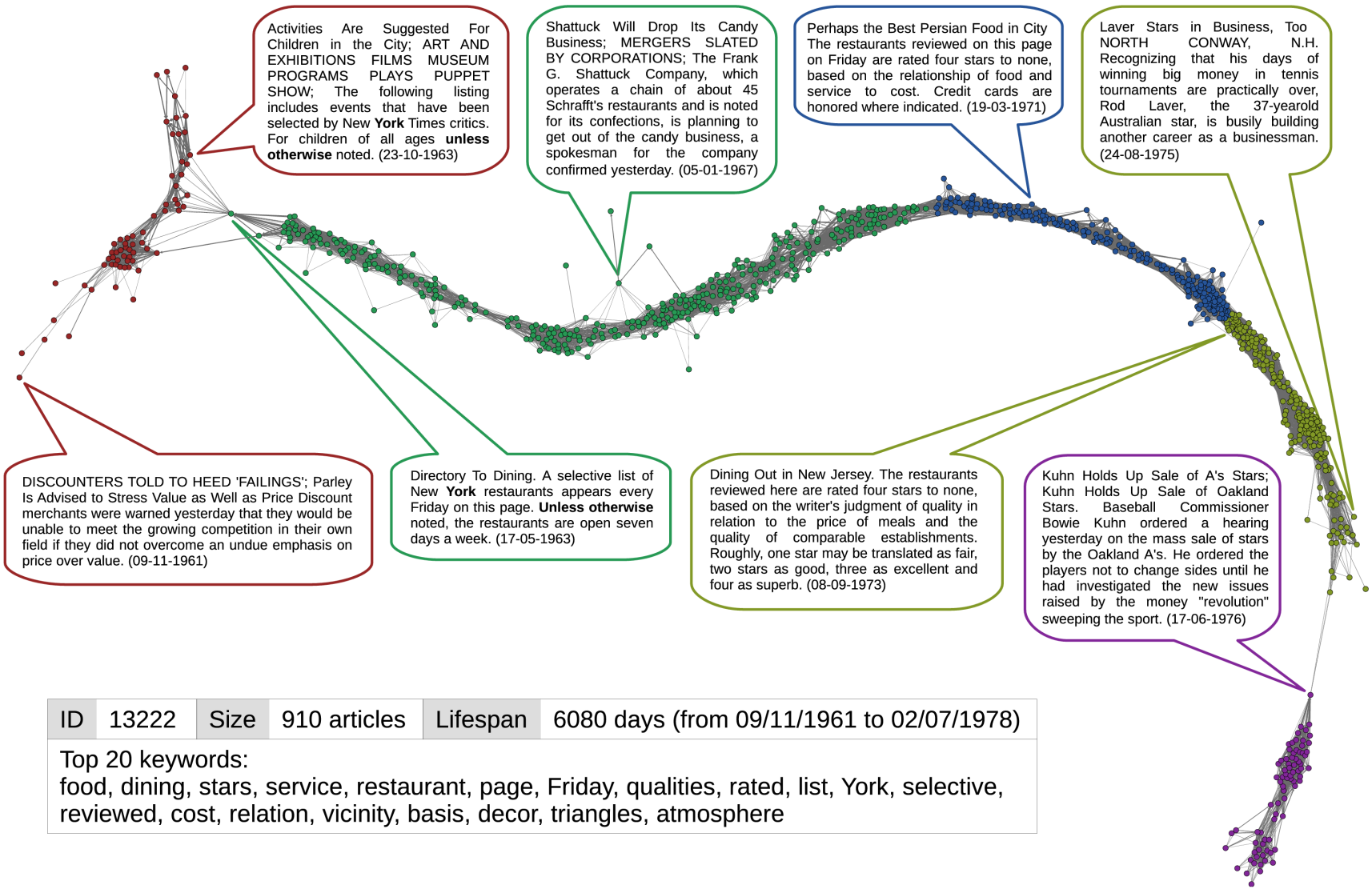
Charakterystyka sieci

1. Liczba węzłów = **6.45 mln** artykułów
2. Liczba linków = **201.54 mln**
3. Średni stopień = **62.5**
4. Liczba komponentów = **23114**
5. Quasi-wykładniczy rozkład podobieństw
6. Quasi-potęgowy rozkład odstępów czas.
7. Szerokie rozkłady stopni wierzchołków
8. Szeroki rozkład wielkości komponentów





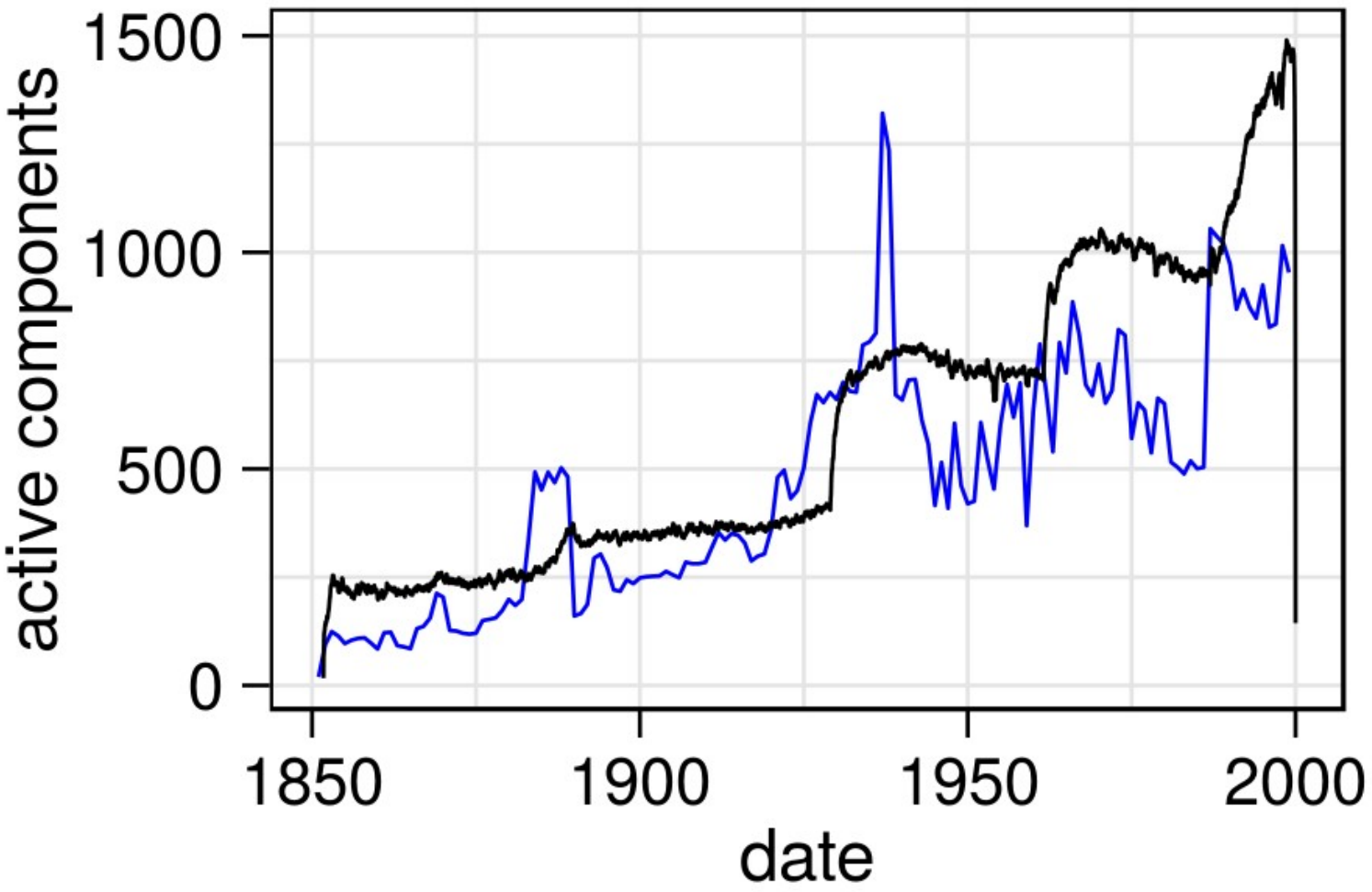
Komponenty



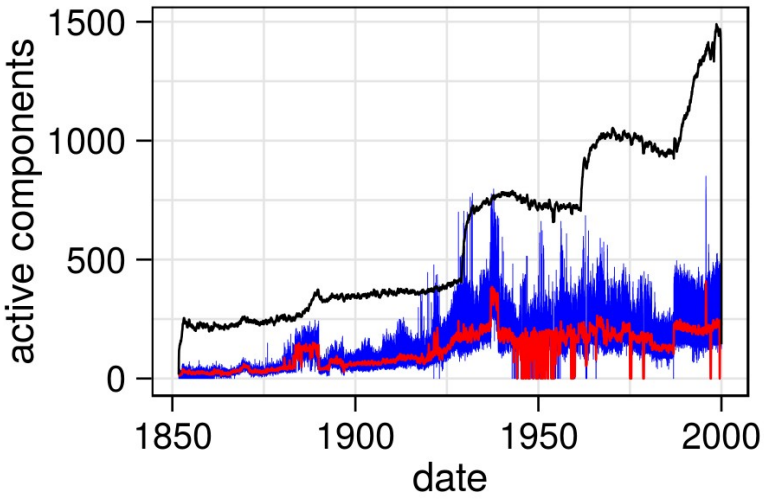
ID	13222	Size	910 articles	Lifespan	6080 days (from 09/11/1961 to 02/07/1978)

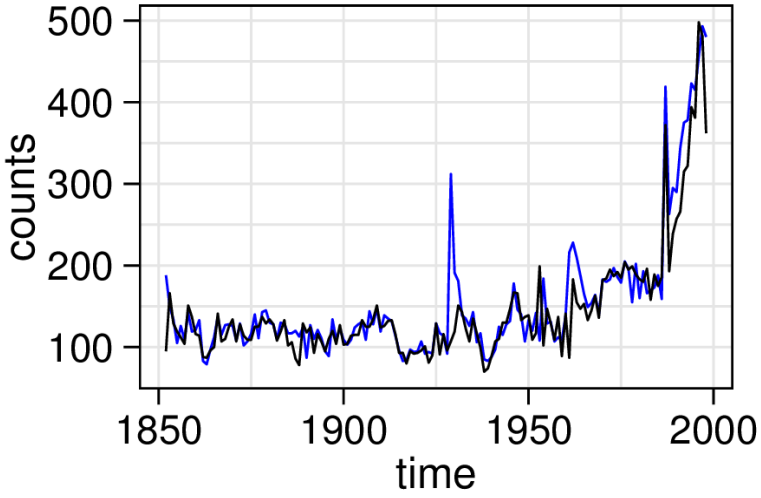
Top 20 keywords:
food, dining, stars, service, restaurant, page, Friday, qualities, rated, list, York, selective, reviewed, cost, relation, vicinity, basis, decor, triangles, atmosphere

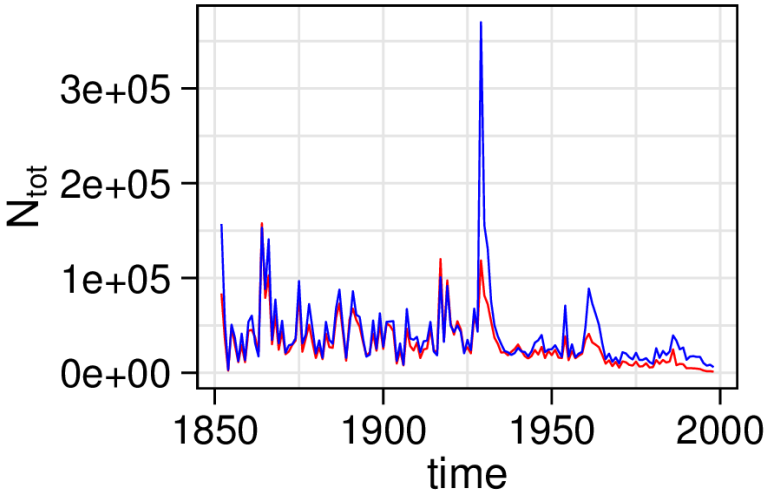
Komponenty

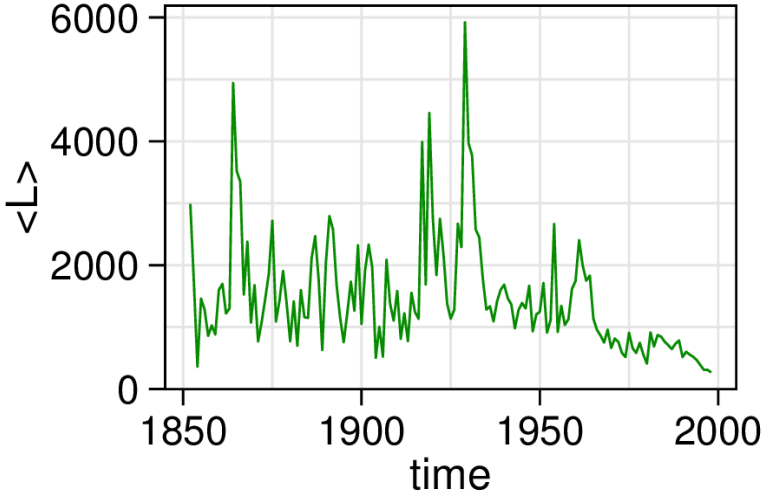


Komponenty


active components
1500
1000
500
0
1850
1900
1950
2000
date


counts
500
400
300
200
100
1850
1900
1950
2000
time


N_{tot}
3e+05
2e+05
1e+05
0e+00
1850
1900
1950
2000
time


<L>
6000
4000
2000
0
1850
1900
1950
2000
time

Możliwe przyczyny

1. Więcej artykułów w gazecie.

2. Zmiana układu gazety – np. więcej działów.

3. Ważne wydarzenia historyczne

    a) 24.10.1929 – Czarny czwartek – krach na giełdzie nowojorskiej

    b) 25.07.1961 – przemówienie J.F. Kennedy na temat kryzysu Berlińskiego

    c) 30.10.1961 – Sowieci detonują testową bombę wodorową

    d) 18.11.1961 – JFK wysyła 18 tys. doradców wojskowych do Wietnamu

4. Drastyczne zmiany językowe.

5. Inne, trudne do uchwycenia zmiany, m.in. zwiększenie liczby dziennikarzy, tematów poruszanych na łamach gazety.

Coś wspólnego

1920 – pierwsze komercyjne stacje radiowe w USA

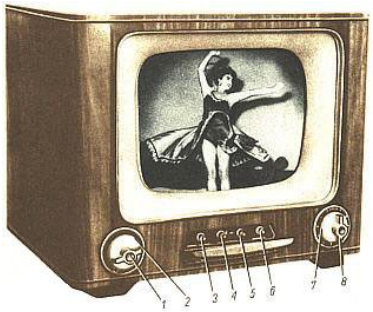
1927 – powstanie Federalnej Komisji Radiowej

1930 – gazety rozpoczynają wojnę przeciwko radiu

1960 – liczba gospodarstw domowych posiadających telewizor osiąga 80%

1960 – seria debat Nixon-Kennedy bije rekordy popularności (70 mln widzów)

25.04.1961 – Orvil R. Dreyfoos zastępuje swojego teścia Arthura Hays Sulzbergera na stanowisku właściciela I wydawcy The New York Times



Podsumowanie

1. Nietrywialna struktura sieci artykułów z NYT
2. Wiele długich komponentów pomimo “szumu”
3. Zaskakujące skoki w liczbie aktywnych komponentów w latach 1929 i 1961
4. Wyraźnie widoczne momenty wzrostu konkurencyjności nowych mediów

DZIĘKUJĘ ZA UWAGĘ