

# Transparency and Topic Models

**Hanna Wallach**  
(AKA “Logistic Aggression”)

wallach@cs.umass.edu  
<http://www.cs.umass.edu/~wallach/>

# Transparency in the US

[ISOO, 2011]

**UNCLASSIFIED**

Approved for Release  
Date 23 JUL 1987

OUTGOING TELEGRAM Department of State

INDICATE:  COLLECT  
 CHARGE TO

Q1

SANITIZED

[REDACTED]  
TITLE: INTERNATIONAL CONGRESS OF SPACE MEDICINE [REDACTED]

DCCREFERENCE: [REDACTED] 6876.

INFLCCATION: 29 JAN 76, 2 PP [REDACTED]  
MEXICO

INFCOATE: [REDACTED] 7509.  
[REDACTED]

- 52.8 million pages reviewed for declassification
- 26.7 million pages declassified
- \$11.36 billion spent on administration of the US government classification system

# Declassified Documents

[Gale, 2012]

~~SECRET~~ NO FOREIGN DISSEM

CENTRAL INTELLIGENCE AGENCY  
WASHINGTON, D.C. 20505

29 January 1968

MEMORANDUM FOR: The Honorable Walt W. Rostow  
Special Assistant to the President  
The White House

SUBJECT : Coal and Electric Power Shortages  
in Communist China

1. Al Jenkins asked that we prepare the attached memorandum on shortages of coal and electric power in Communist China for your information. We have also included excerpts from individual reports of shortages to give you some feeling for the information available.

2. While there is no question that the shortages are widespread, it is extremely difficult to quantify the decline in industrial output caused by these shortages or by other effects of the Cultural Revolution.

*Edward W. Proctor*

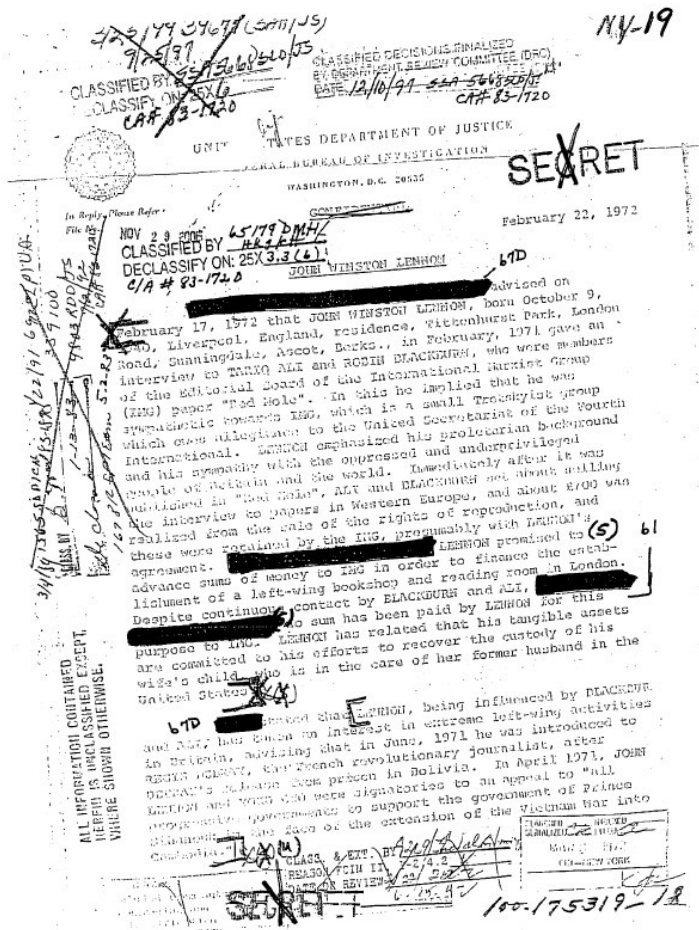
EDWARD W. PROCTOR  
Acting Deputy Director for Intelligence

Attachment:  
Subject Report

DECLASSIFIED  
E.O. 12958, Sec. 3.6  
NLJ 92-193  
By Cb, NARA Date 10-31-97

- Date issued
- Date declassified
- Document type
- Source institution
- Classification level
- Document text

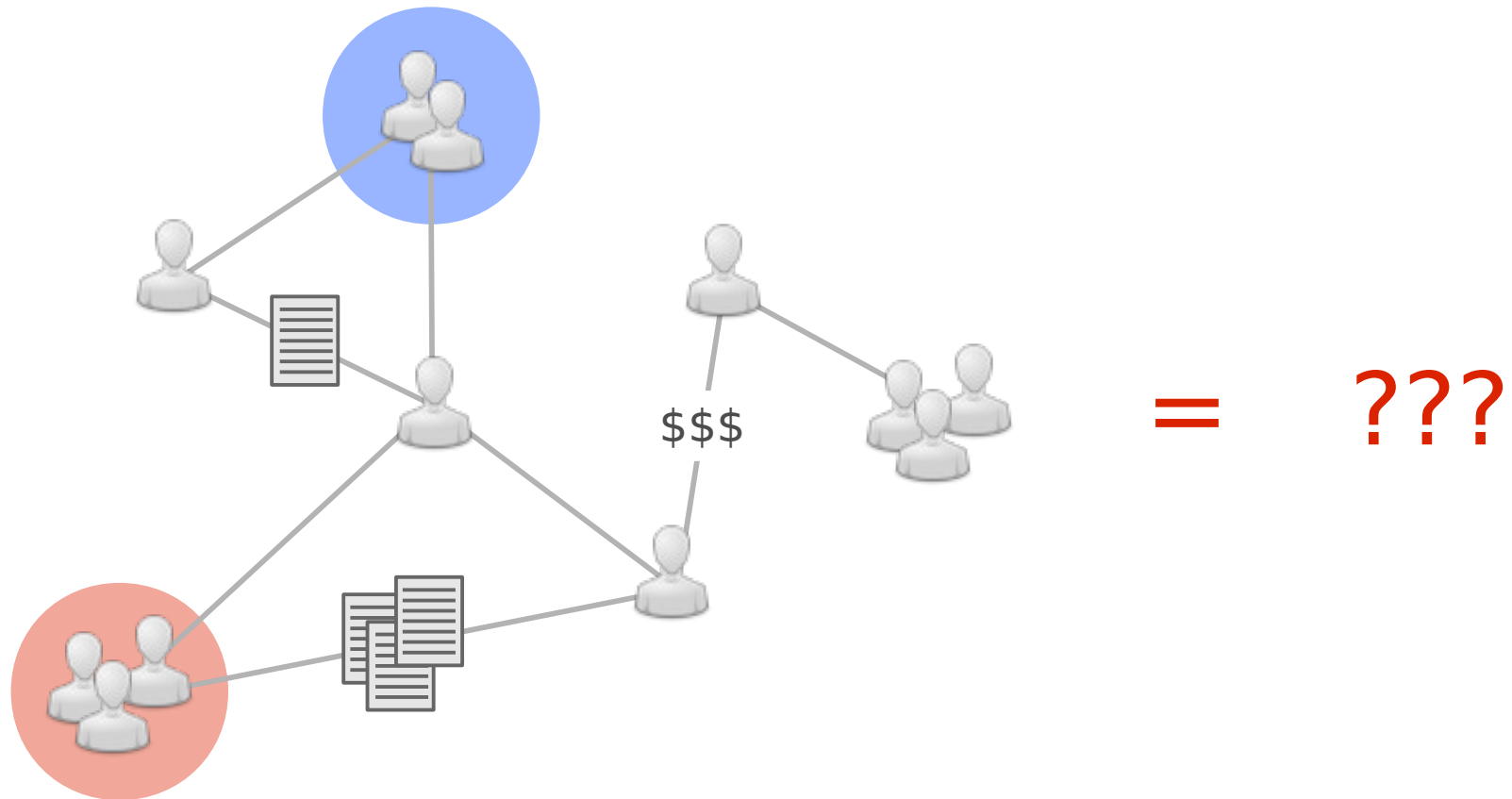
# Text Tells All



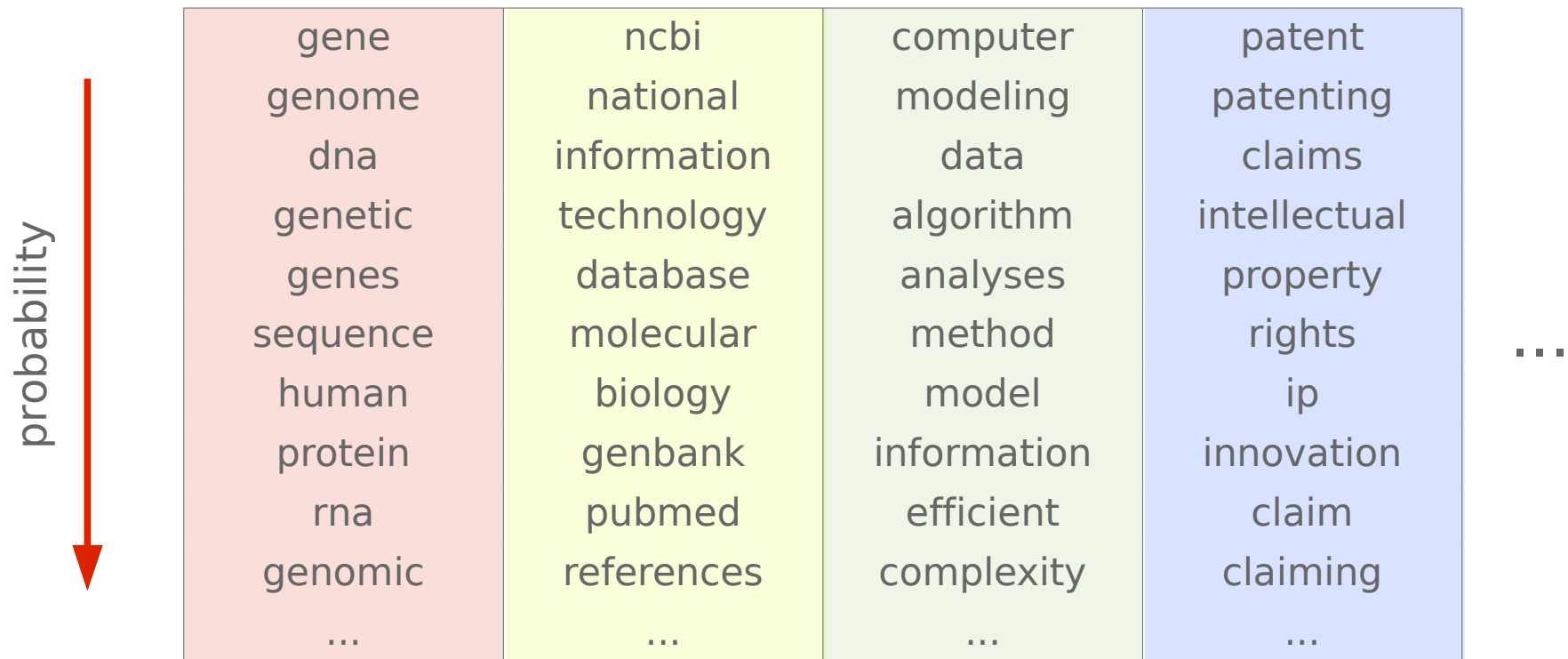
“The FBI has released a new cache of material on John Lennon. The file contains little, if any, new information about Lennon, though it does present some bizarre details, like a description of an antiwar activist trying to **train a parrot to speak profanities..”**

— NYT, 25 September 2007

# Exploratory Analysis



# Topics and Words



# Documents and Topics

## POLICY FORUM

INTELLECTUAL PROPERTY

### Intellectual Property Landscape of the Human Genome

Kyle Jensen and Fiona Murray\*

Gene patents are the subject of considerable debate and yet, like the term “gene” itself, the definition of what constitutes a gene patent is fuzzy (1). Nonetheless, gene patents that seem to cause the most controversy are those claiming human protein-encoding nucleotide sequences. This category is the subject of our analysis of the patent landscape of the human genome (2).

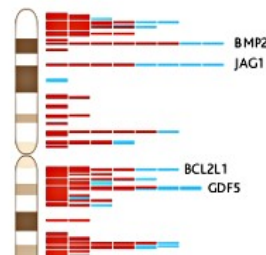
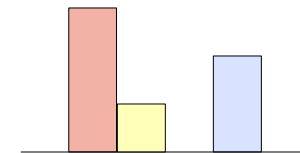
Critics describe the growth in gene sequence patents as an intellectual property (IP) “land grab” over a finite number of human genes (3, 4). They suggest that overly broad patents might block follow-on research (5). Alternatively, gene IP rights may become highly fragmented and cause an anticommens effect, imposing high costs on future innovators and underuse of genomic resources (6). Both situations, critics argue, would increase the costs of genetic diagnostics, slow the development of new medicines, stifle academic research,

tinguishing patents on the human genome from those on other species (23).

Our detailed map was developed using bioinformatics methods to compare nucleotide sequences claimed in U.S. patents to the human genome. Specifically, this map is based on a BLAST (24) homology search linking nucleotide sequences disclosed and claimed in granted U.S. utility patents to the set of protein-encoding messenger RNA transcripts contained in the National Center for Biotechnology Information (NCBI) RefSeq (25) and Gene (26) databases. This method allows us to map gene-oriented IP rights to specific physical loci on the human genome (27) (see figure, right). Our approach is highly specific in its identification of patents that actually claim human nucleotide sequences. However, by limiting the search to patents using the canoni-

California, Isis Pharmaceuticals, the former SmithKline Beecham, and Human Genome Sciences. The top patent assignee is Incyte Pharmaceuticals/Incyte Genomics, whose IP rights cover 2000 human genes, mainly for use as probes on DNA microarrays.

Although large expanses of the genome are unpatented, some genes have up to 20 patents asserting rights to various gene uses and manifestations including diagnostic uses, single nucleotide polymorphisms (SNPs), cell lines, and constructs containing the gene. The distribution of gene patents was nonuniform (see figure, page 240, top right): Specific regions of the genome are “hot spots” of heavy patent activity, usually with a one-gene-many-patents scenario (see figure, below). Although less common, there were cases in which a single patent claims many genes, typically as complementary DNA probes used on a microarray (see figure, p. 240, bottom).



# Latent Dirichlet Allocation

[Blei, Ng & Jordan, '03]



## POLICY FORUM

### INTELLECTUAL PROPERTY

#### Intellectual Property Landscape of the Human Genome

Kyle Jensen and Fiona Murray\*

Gene patents are the subject of considerable debate and yet, like the term "gene" itself, the definition of what constitutes a gene patent is fuzzy (1). Nonetheless, gene patents that seem to cause the most controversy are those claiming human protein-encoding nucleotide sequences. This category is the subject of our analysis of the patent landscape of the human genome (2). Critics describe the growth in gene sequence patents as an intellectual property (IP) "land grab" over a finite number of human genes (3, 4). They suggest that overly broad patents might block follow-on research (5). Alternatively, gene IP rights may become highly fragmented and cause an anticommons effect, imposing high costs on future innovators and underuse of genomic resources (6). Both situations, critics argue, would increase the costs of genetic diagnostics, slow the development of new medicines, stifle academic research, distinguishing patents on the human genome from those on other species (23).

Our detailed map was developed using bioinformatics methods to compare nucleotide sequences claimed in U.S. patents to the human genome. Specifically, this map is based on a BLAST (24) homology search linking nucleotide sequences disclosed and claimed in granted U.S. utility patents to the set of protein-encoding messenger RNA transcripts contained in the National Center for Biotechnology Information (NCBI) RefSeq (25) and Gene (26) databases. This method allows us to map gene-oriented IP rights to specific physical loci on the human genome (27) (see Figure, right). Our approach is highly specific in its identification of patents that actually claim human nucleotide sequences. However, by limiting the search to the canonical

California, Isis Pharmaceuticals, the former SmithKline Beecham, and Human Genome Sciences. The top patent assignee is Incyte Pharmaceuticals/Incyte Genomics, whose IP rights cover 2000 human genes, mainly for use as probes on DNA microarrays.

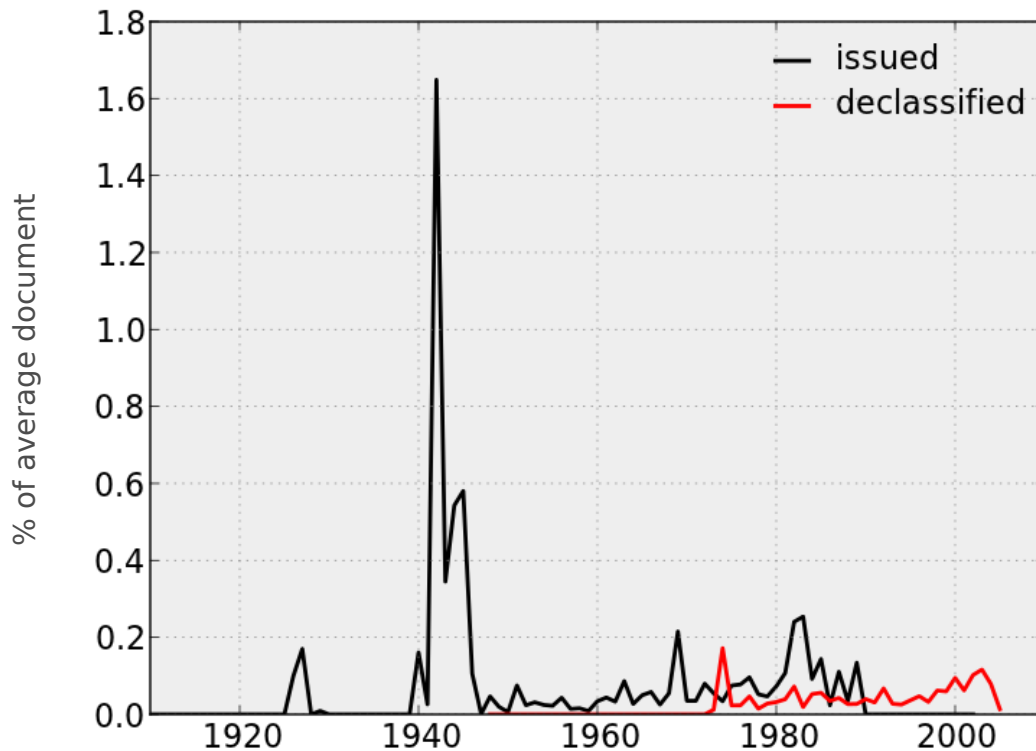
Although large expanses of the genome are unpatented, some genes have up to 20 patents asserting rights to various gene uses and manifestations including diagnostic uses, single nucleotide polymorphisms (SNPs), cell lines, and constructs containing the gene. The distribution of gene patents was nonuniform (see figure, page 240, top right): Specific regions of the genome are "hot spots" of heavy patent activity, usually with a one-gene-many-patents scenario (see figure, below). Although less common, there were cases in which a single patent claims many genes, typically as complementary DNA probes used on a microarray (see figure, p. 240, bottom).

?



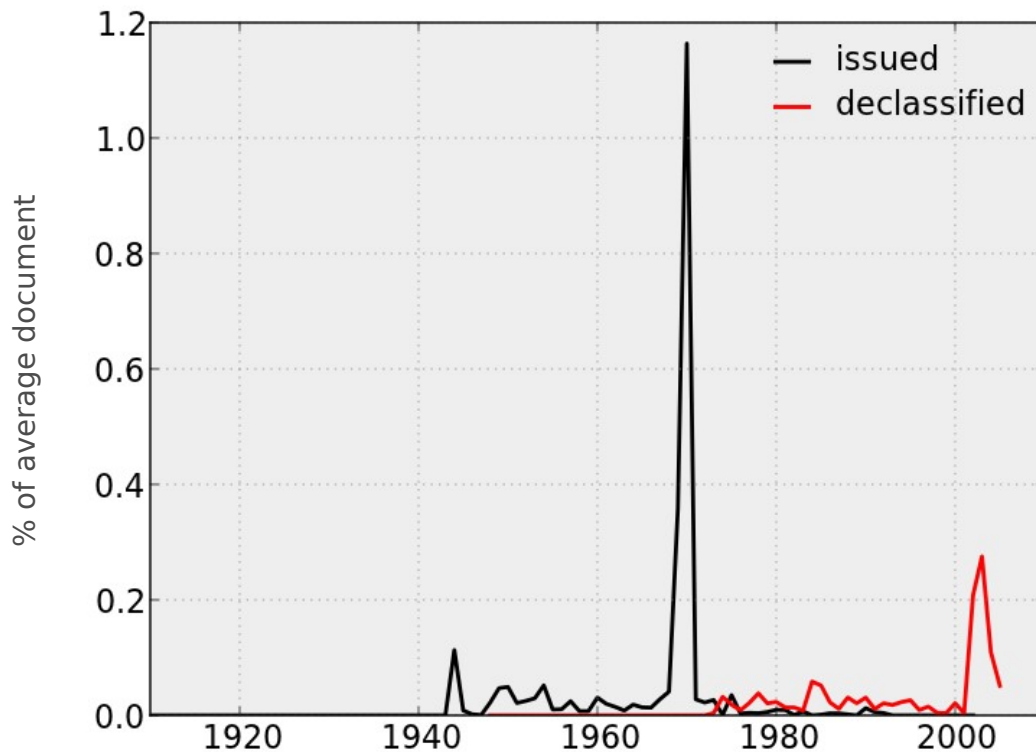
# Inferred Topics

church  
catholic  
pope  
vatican  
religious  
bishop  
cardinal  
archbishop  
priests  
paul  
...



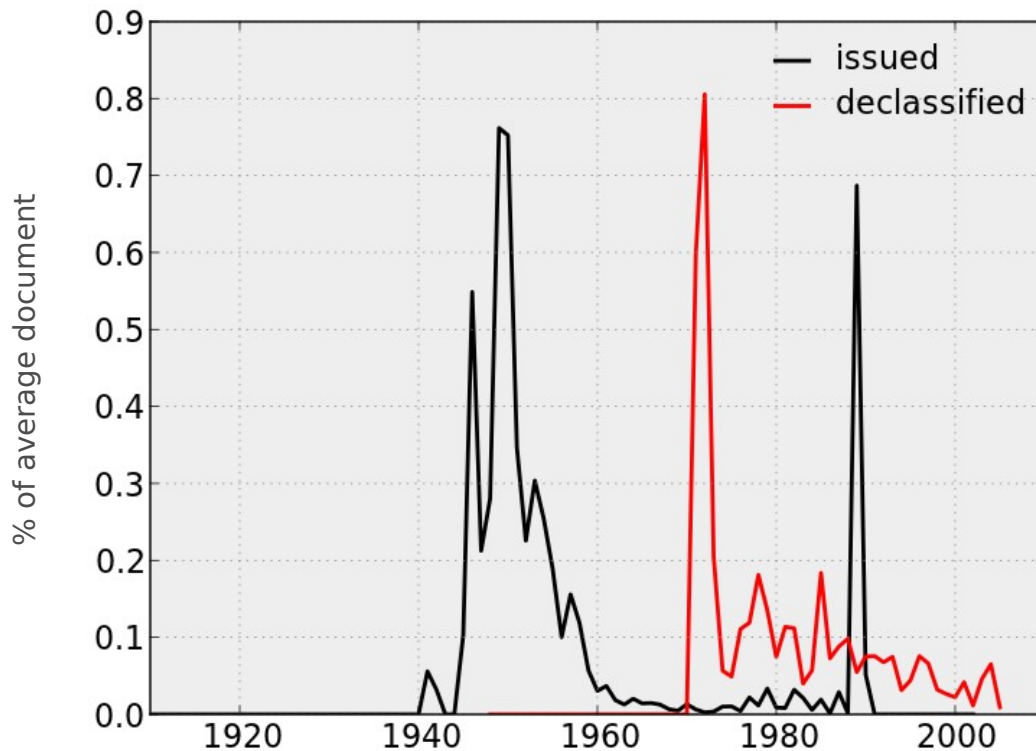
# Inferred Topics

draft  
service  
manpower  
volunteer  
selective  
age  
calls  
volunteers  
deferments  
pay  
...

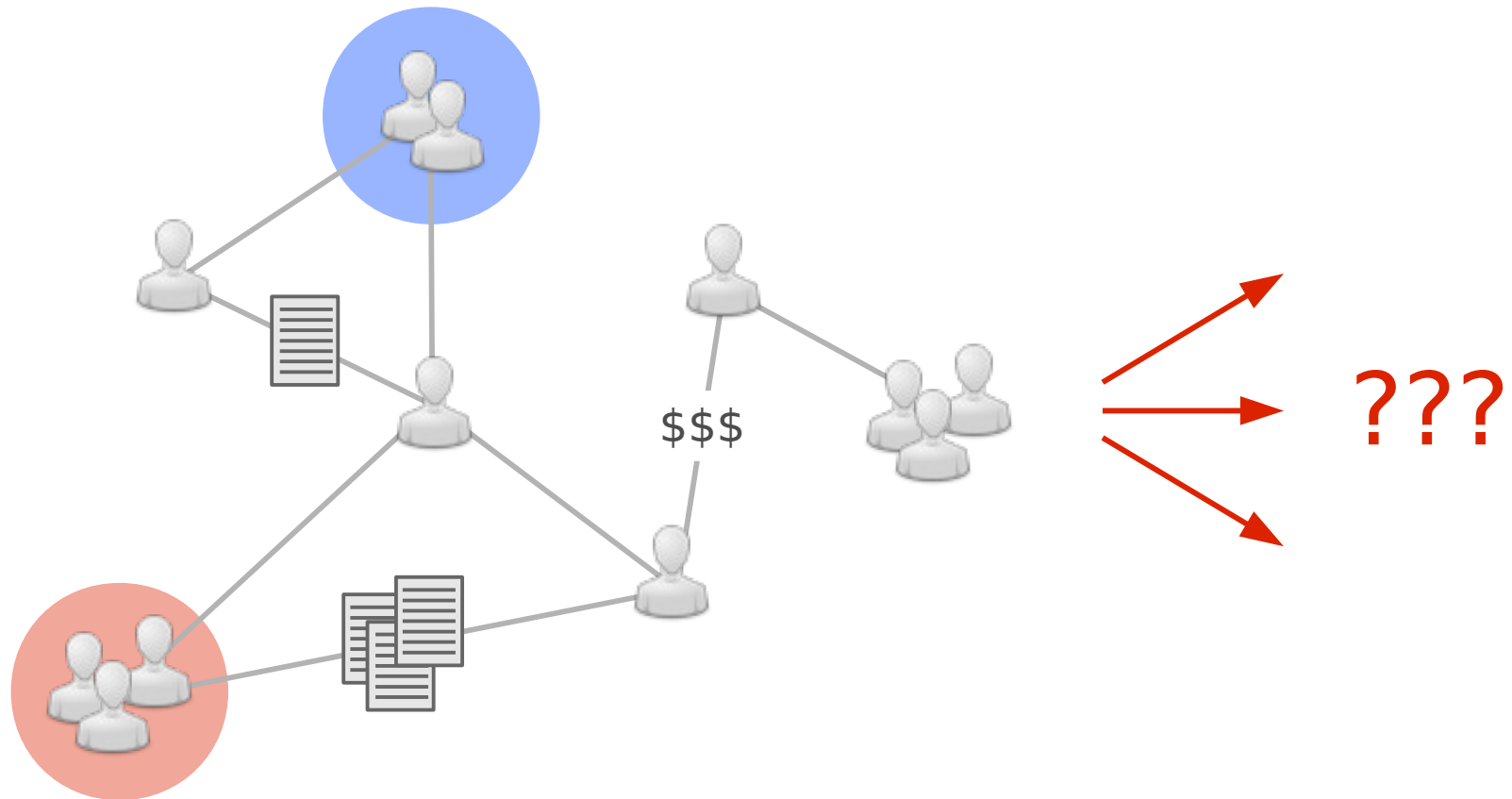


# Inferred Topics

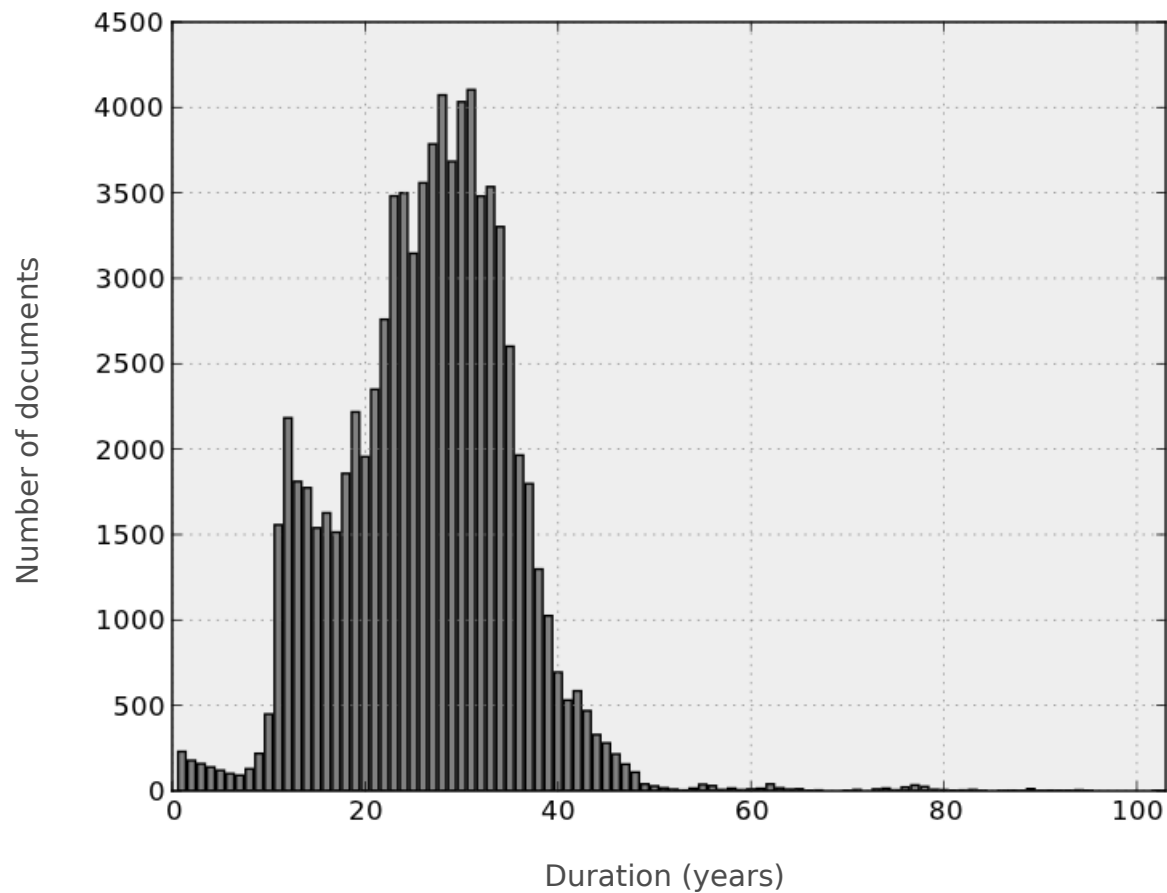
atomic  
weapon  
bomb  
bombs  
weapon  
energy  
thermonuclear  
development  
hydrogen  
stockpile  
...



# Predictive Analysis



# Classification Durations



# Duration and Content

HIS APPROACH WAS, "WELL, OF COURSE, WE KNOW THERE ISN'T ANYTHING TO THIS ALLEGED PHENOMENON (FLYING SAUCERS), BUT ON THE OTHER HAND". DURING HIS TALK SHKLOVSKIY AND OTHER SOVIETS JOKED AND LAUGHED AND OBVIOUSLY DID NOT TAKE THE SPEAKER'S REMARKS SERIOUSLY.

  
14 years

57 years

  
CENTRAL INTELLIGENCE GROUP

SOVIET CAPABILITIES FOR THE DEVELOPMENT AND PRODUCTION  
OF CERTAIN TYPES OF WEAPONS AND EQUIPMENT

1. Herein is presented an estimate of Soviet capabilities in the development and production, during the next ten years, of certain weapons and equipment, as follows:

# Thanks!

Acknowledgements: Bruce Desmarais, Rachel Shorey, Mr. Sprinkles

