PREVIEW OF TEXT ANALYSIS

· *Jesse Lecy* ·

Text as Data

1. pre-processing
2. tokenization
3. feature selection
4. modeling

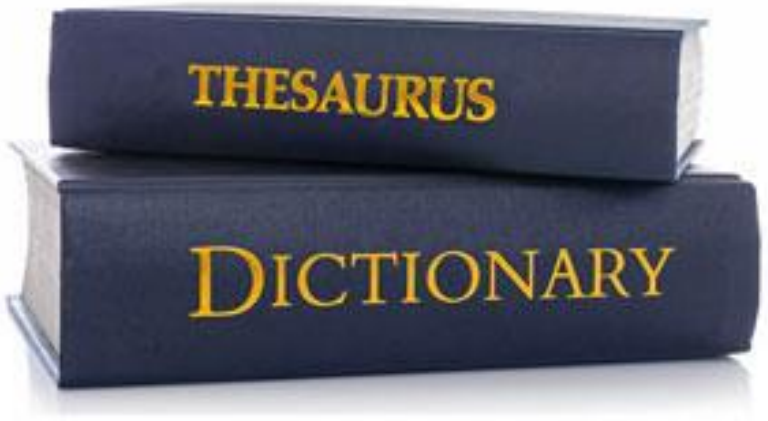
THE CORPORATION'S SPECIFIC PURPOSE IS TO SUPPORTS AFFORDABLE HOUSING, COMMUNITY DEVELOPMENT AND ECONOMIC DEVELOPMENT OF THE CITY AND COUNTY OF SAN FRANCISCO'S ECONOMICALLY DISADVANTAGED INDIVIDUALS AND COMMUNITIES, BY LENDING TO, INVESTING IN, AND DIRECTLY ACQUIRING SUCH AFFORDABLE HOUSING AND RELATED COMMUNITY DEVELOPMENT REAL ESTATE ASSETS.

~~the~~ corporation specific purpose ~~is to~~ support AFFORDABLE_HOUSING, community development ~~and~~ ECONOMIC_DEVELOPMENT ~~of the~~ city and county of SAN_FRANCISCO economically disadvantaged individuals and communities by lending ~~to~~ investing ~~in and~~ directly acquiring ~~such~~ AFFORDABLE_HOUSING ~~and~~ related community development REAL_ESTATE assets

1. Remove punctuation
2. Delete words with little information value (“stop words” in quanteda)
3. Identify compound constructs (apply “dictionary”)

STEMMING

LEND	RELATE
LEND~~ing~~	RELATE~~d~~



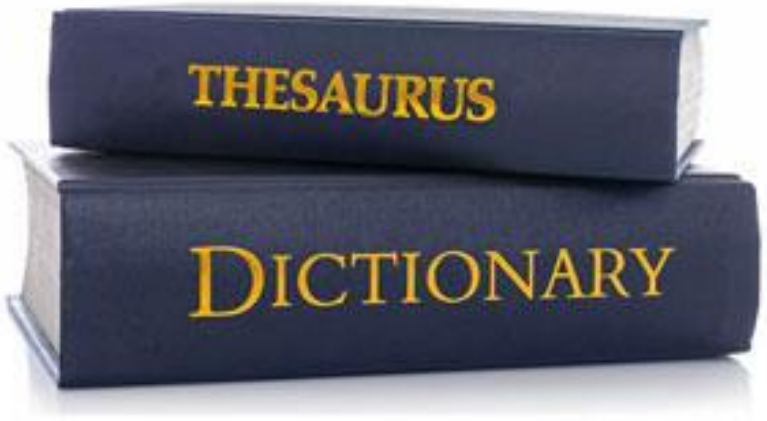
DISAMBIGUATION

George W. Bush

George Bush Jr.

President Bush

→ GW_BUSH



DISAMBIGUATION

New York City → PLACE

THING ← New York University

New Yorker → STATE OF MIND

STATE ← New York

DOCUMENT FREQUENCY MATRIX (DFM): final output of pre-processing steps in quanteda

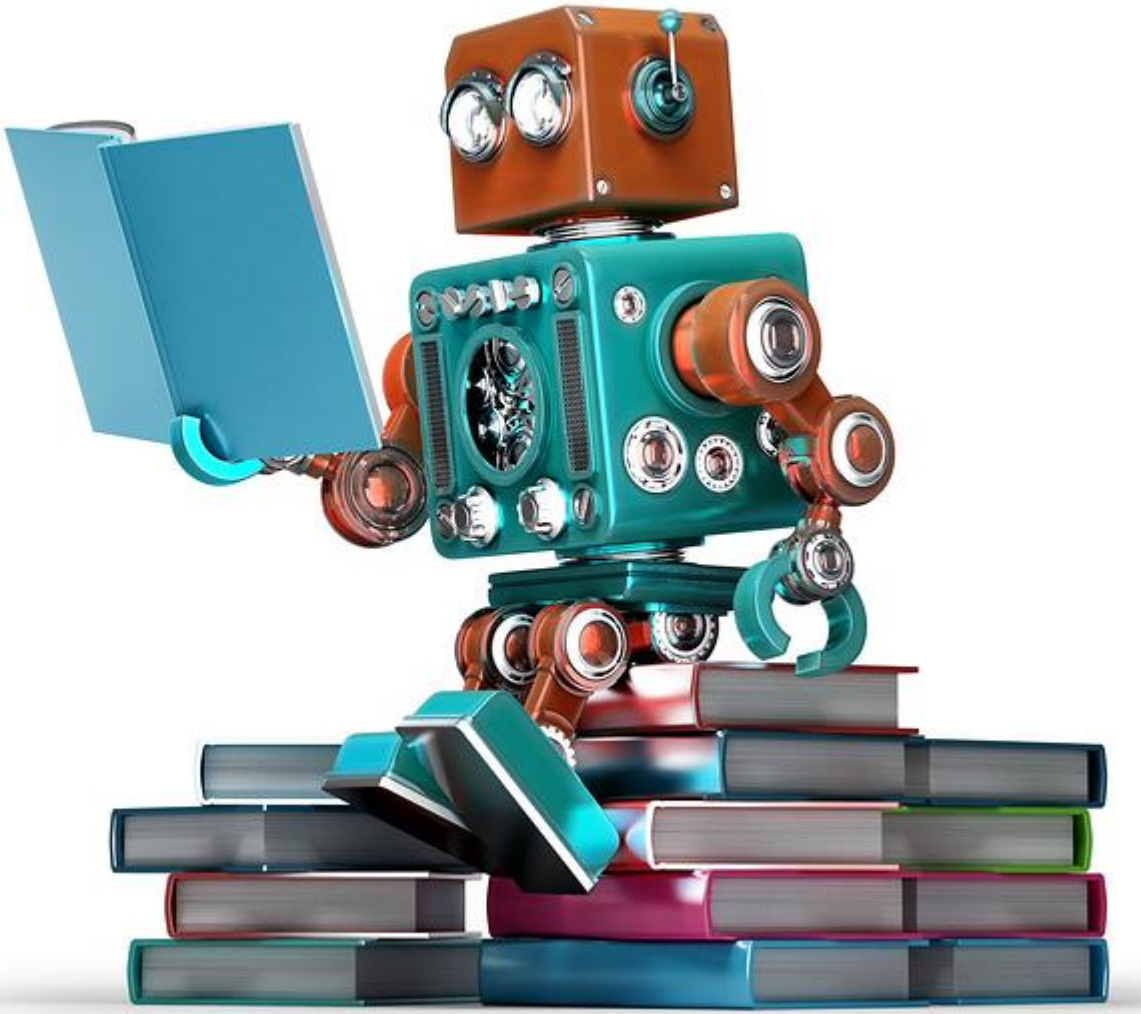
Terms	Documents													
	M 1	M 2	M 3	M 4	M 5	M 6	M 7	M 8	M 9	M 10	M 11	M 12	M 13	M 14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0

Machine Learning

(classification)

7EXT A5 4ATA:

Text is a useful input for many machine learning models for prediction and trend analysis

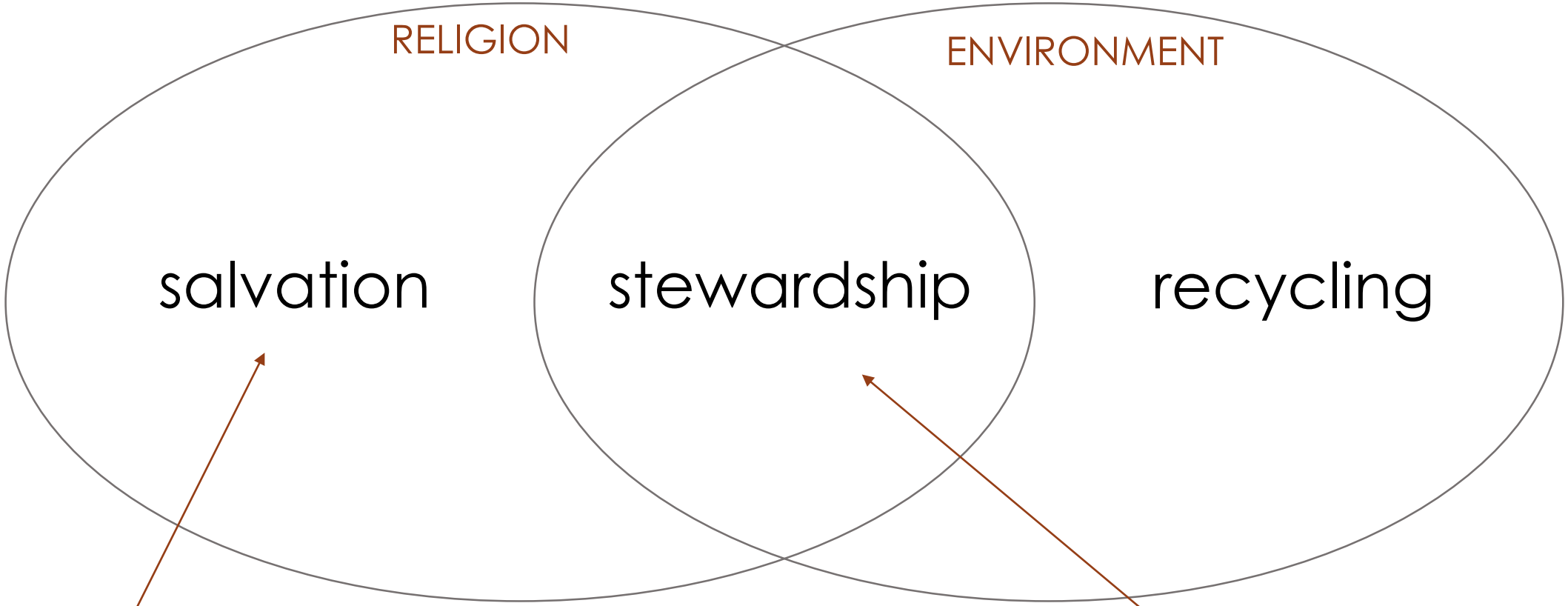


The Netflix Prize

Netflix provided a *training data set* of 100,480,507 ratings that 480,189 users gave to 17,770 movies, graded from 1 to 5 stars.

The *test set* of 1,408,789 ratings is used by the jury to determine potential prize winners.

In 2009 the grand prize of $1,000,000 was given to the BellKor's Pragmatic Chaos team which
improved Netflix's own algorithm by 10%



Distinctive constructs

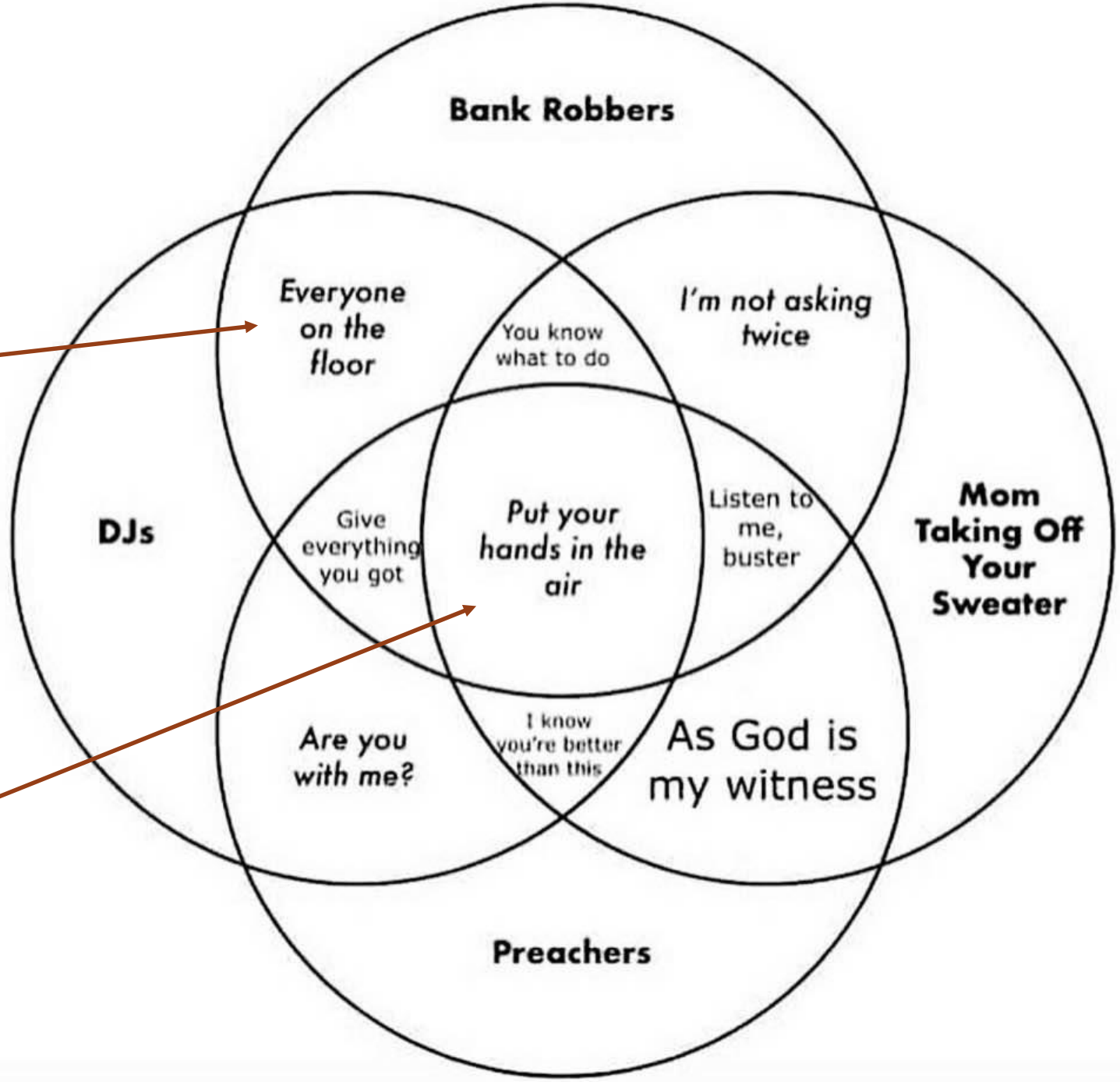
I. Arts, Culture, and Humanities
II. Education
III. Environment and Animals
IV. Health
V. Human Services
VI. International, Foreign Affairs
VII. Public, Societal Benefit
VIII. Religion Related

Shared constructs

Information theory

HIGH Signal
to Noise Ratio
(eliminates half of
possible sources)

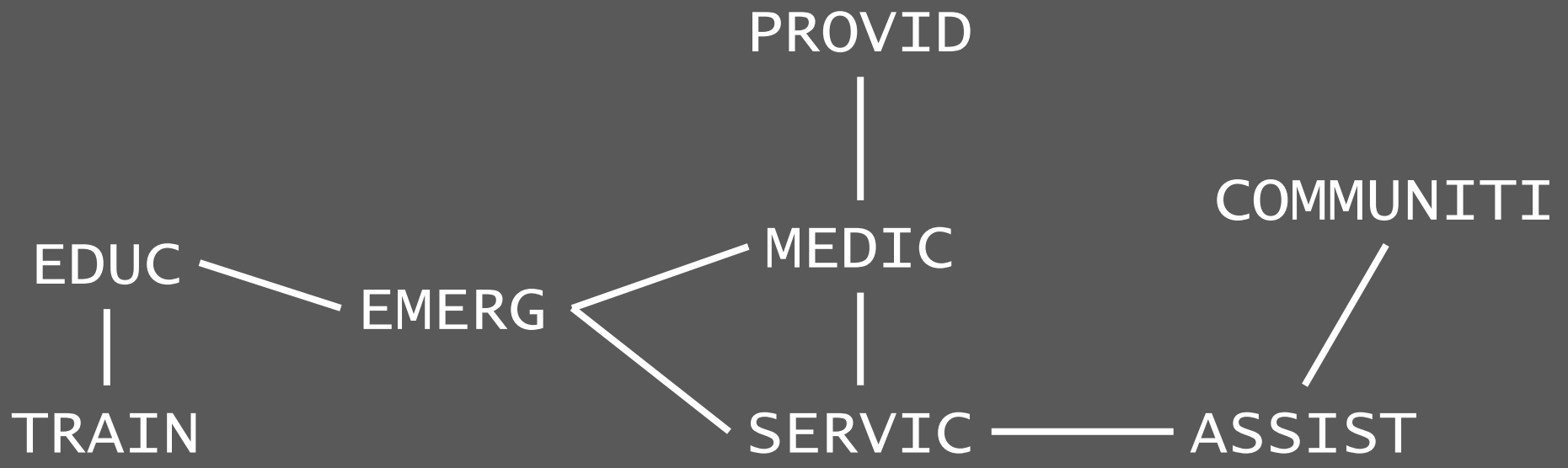
LOW Signal
to Noise Ratio
(does not distinguish
any actors)

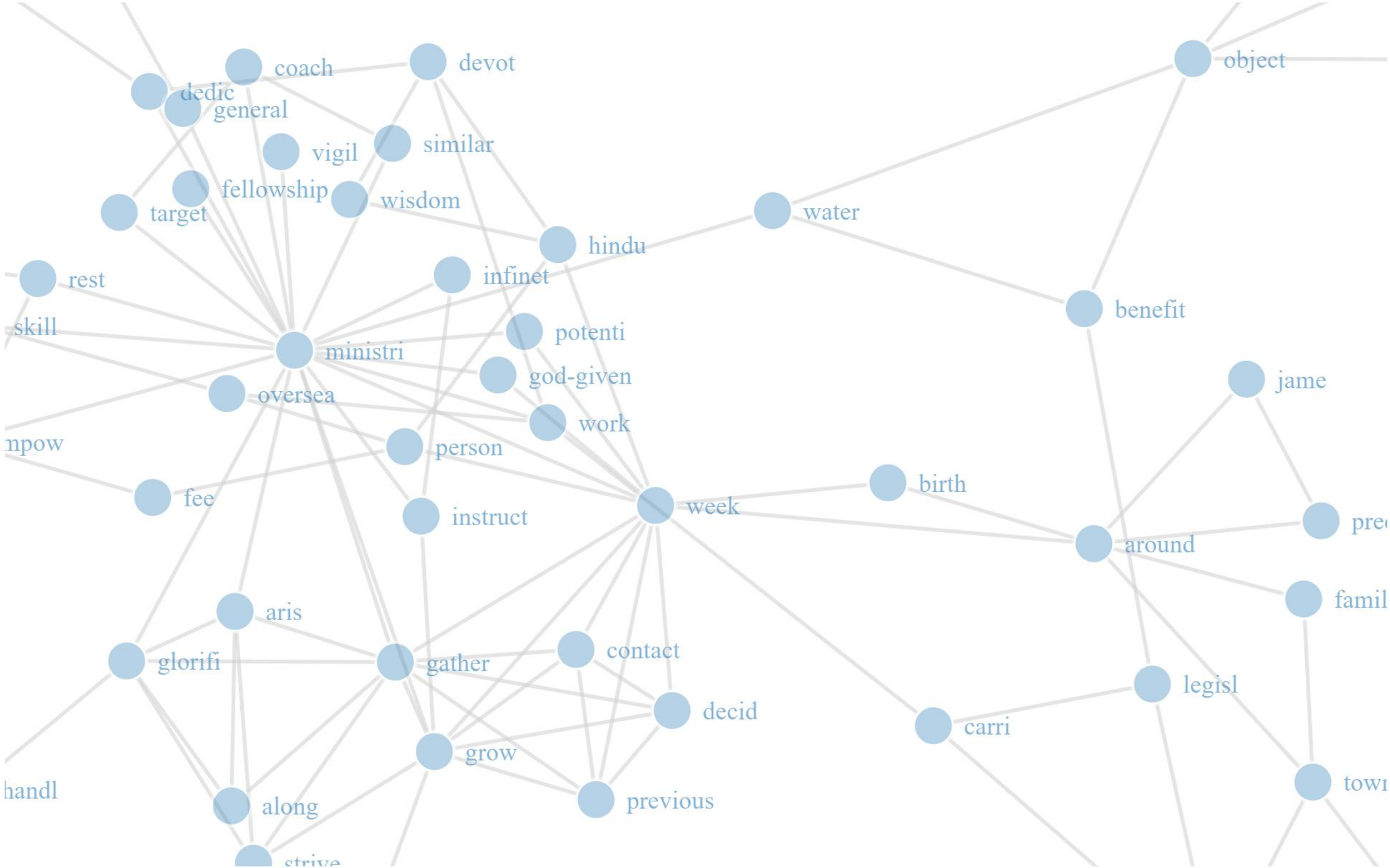


To educate, train and assist in providing emergency medical service for the community.

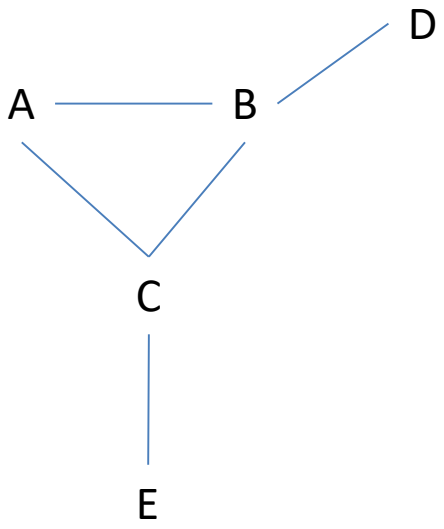
```
"EDUC" "TRAIN" "ASSIST"
"PROVID" "EMERG" "MEDIC"
"SERVIC" "COMMUNITI"
```

Semantic Networks

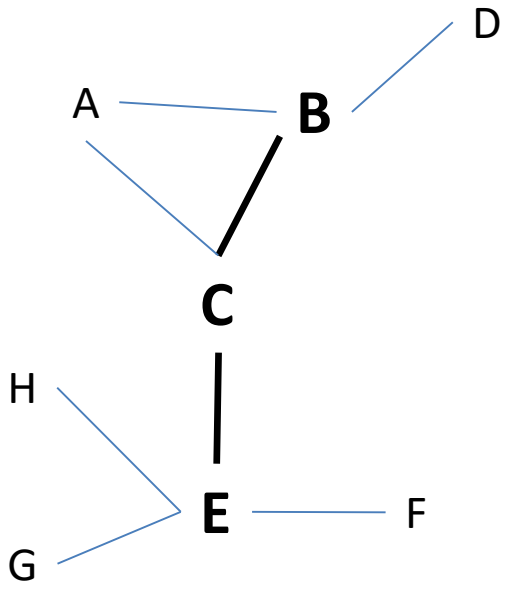



coach
devot
dedic
general
vigil
similar
fellowship
wisdom
target
hindu
rest
infinet
skill
potenti
ministri
god-given
oversea
work
person
fee
instruct
week
aris
glorifi
gather
contact
decid
grow
previous
along
handl
object
water
benefit
jame
birth
around
famil
legisl
carri
towi

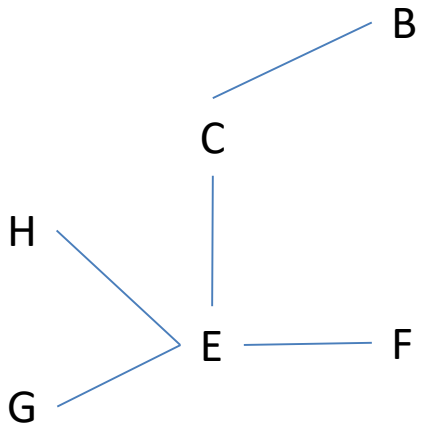
Mission Statement 1



Union (all statement) and **Intersection**



Mission Statement 2

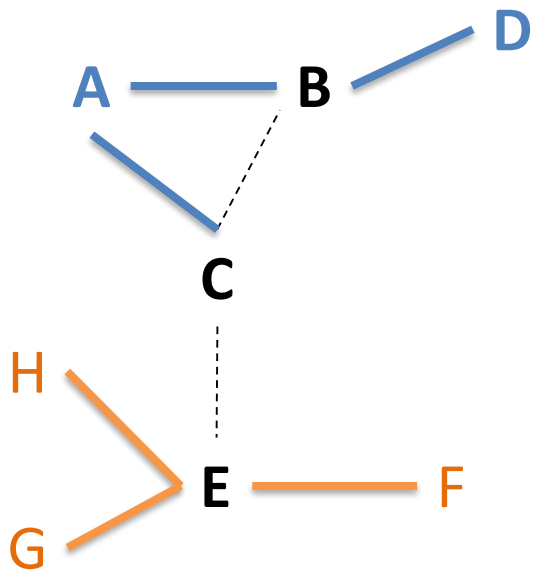


Intersection represents components of mission statements common to all nonprofits.

Analyzing Missions by Types of Nonprofits

Mission Statement Components Unique to Org 1:

A-B
A-C
B-D



H-E
G-E
E-F

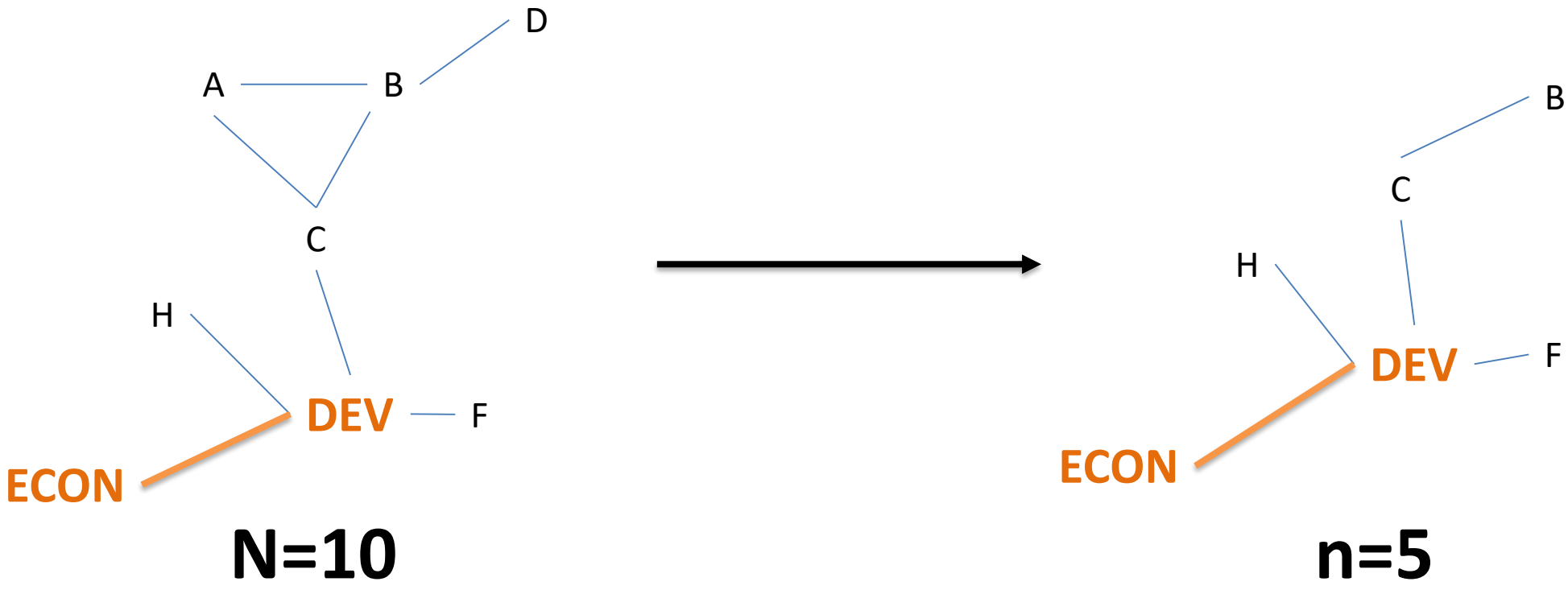
Mission Statement Components Unique to Org 2

Doesn't work well with dense weighted graphs!

Data structure of a weighted network:

Freq ALL	Freq GROUP	Term 1	Term 2
10	5	econ	dev
7	4	self	reliance
5	3	dev	con
5	2	globla	econ
4	2	local	econ
4	1	soc	econ
3	2	econ	socialism
3	3	finance	global
3	2	global	finance
3	2	global	impsm
3	1	impsm	global
3	1	impsm	invasion

Is it significant that **economic development** was mentioned **5 times** by a specific type of organization?



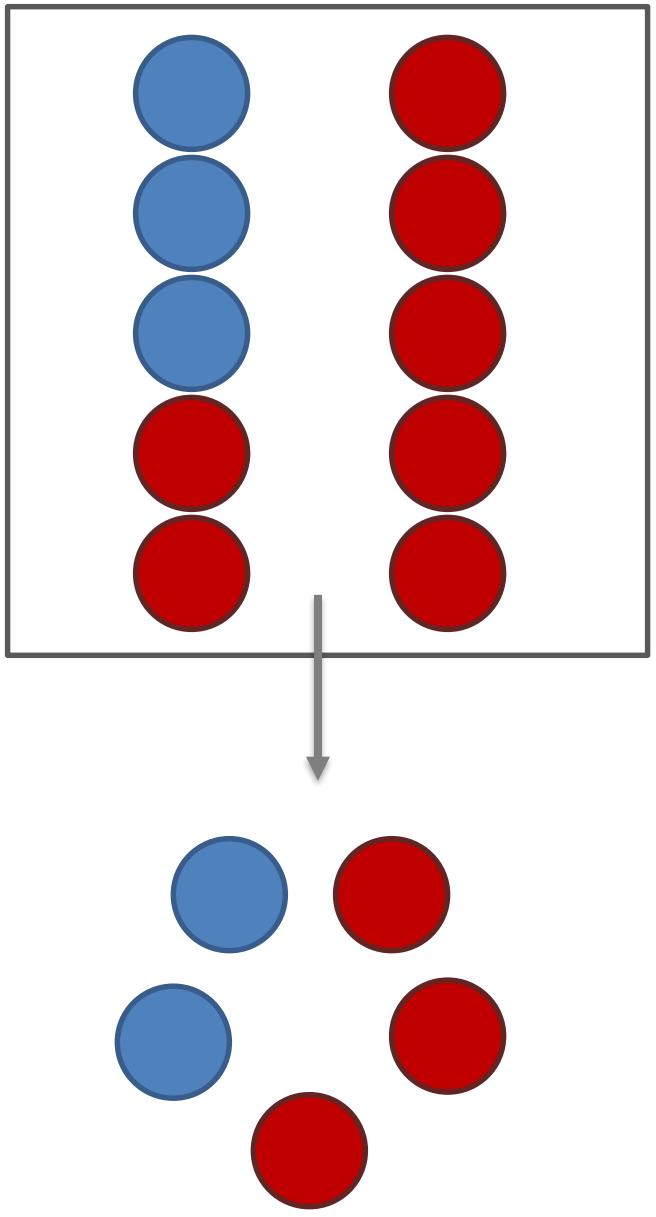
N=10

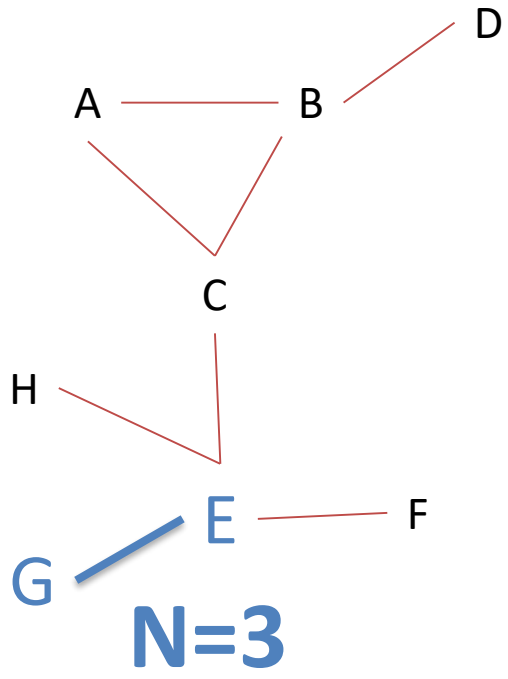
n=5

How often will a **random sample** from the full weighted network produce the **observed number** of "statements" (semantic network ties)?

Is it significant that **Org Statement** significant?

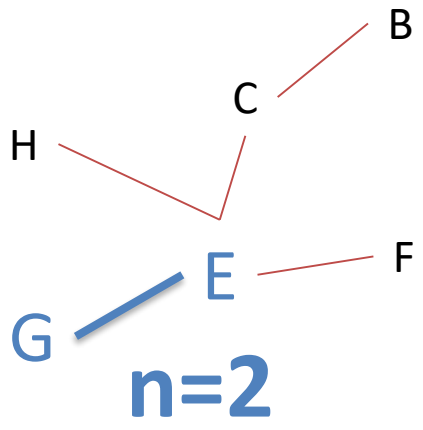
What is the probability of selecting **2 blue balls** from a **sample of 5 balls**?





$$\Pr(\, blue = 2 \mid n = 5\,) = \frac{\binom{3}{2}\binom{7}{3}}{\binom{10}{5}}$$

$$= 0.42$$



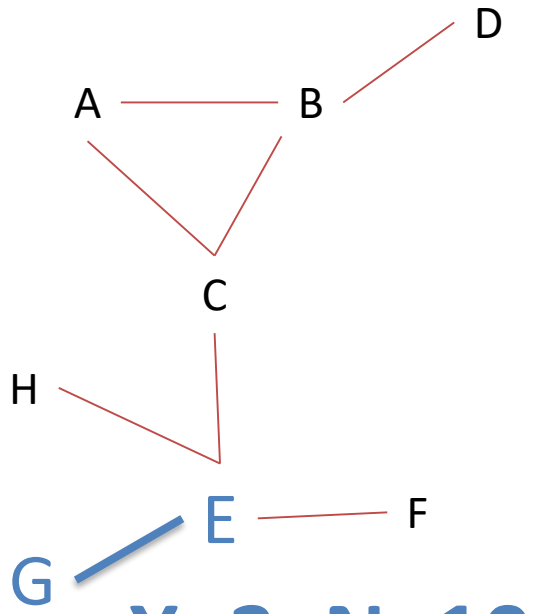
Generalized:

$$\Pr(\mathit{StatementCount} = x \mid \mathit{sample} = k) = \frac{\binom{X}{x}\binom{N-X}{k-x}}{\binom{N}{k}}$$

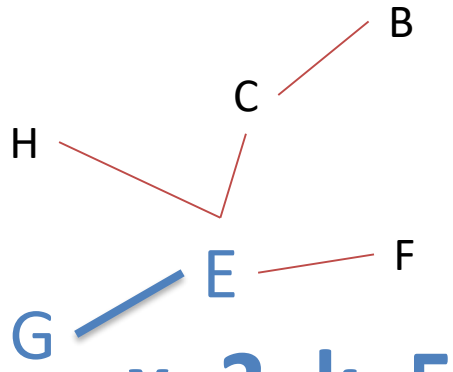
$Where\ X = the\ number\ of\ time\ a\ statement\ appears$

$N = total\ number\ of\ statements$

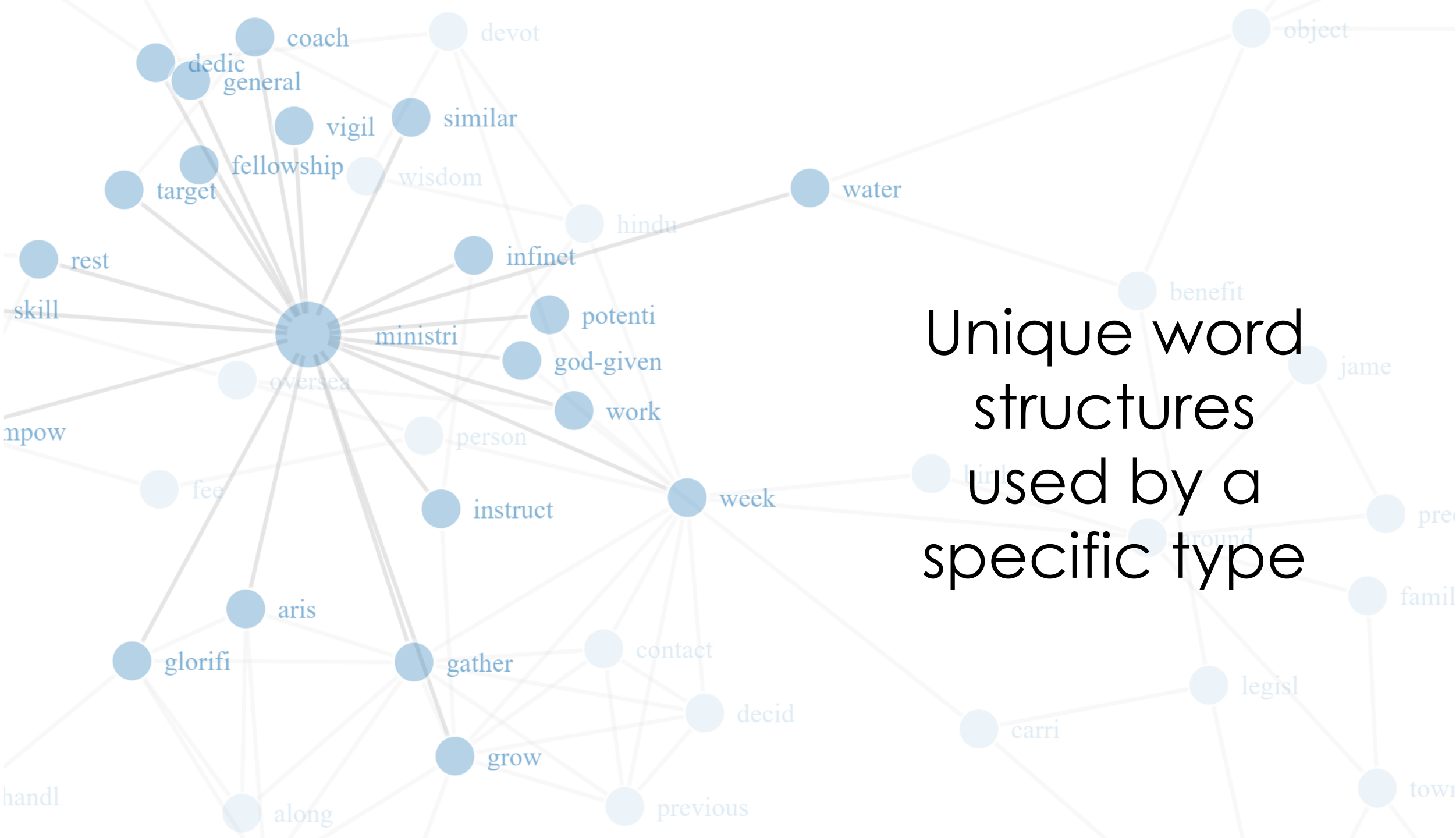
$k = number\ of\ statements\ in\ a\ specific\ period\ or\ group$



X=3, N=10



x=2, k=5



Unique word structures used by a specific type