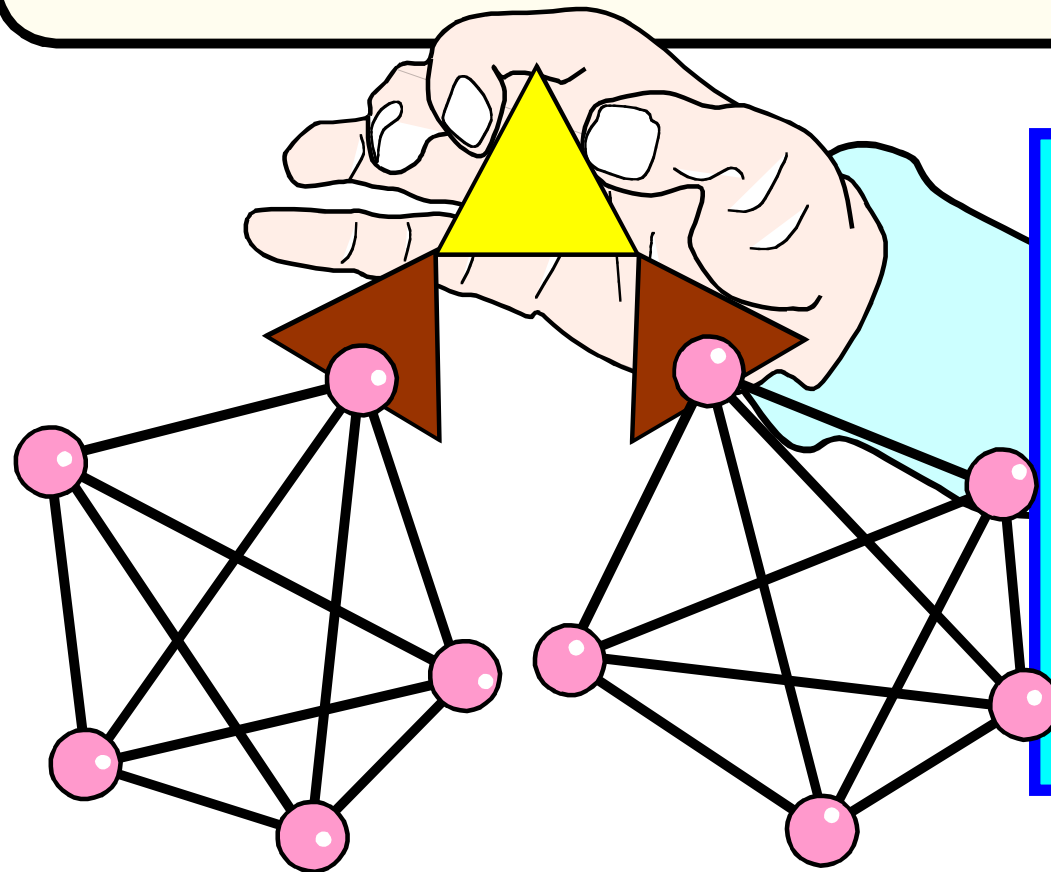


# *Categories, Cliques and Analogies*

## *in Creative Information/Knowledge Management*



**Tony Veale (+ students)**

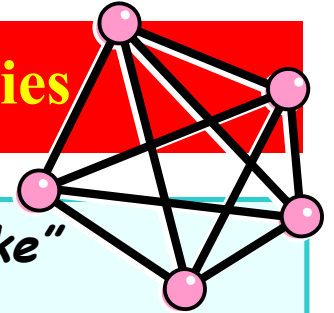
**School of Computer Science  
and informatics**

**University College Dublin**

**[Tony.Veale@UCD.ie](mailto:Tony.Veale@UCD.ie)**

**<http://Afflatus.UCD.ie>**

# Cliques, Categories, Analogies, Metaphors and Ontologies



Ontologies impose structure on a domain by grouping “*like with like*”

Good categories have high intra-category similarity, low inter-category similarity

The members of coherent (*tight-knit*) categories resemble social “cliques”

Members closely connected: new members must bond with all existing members

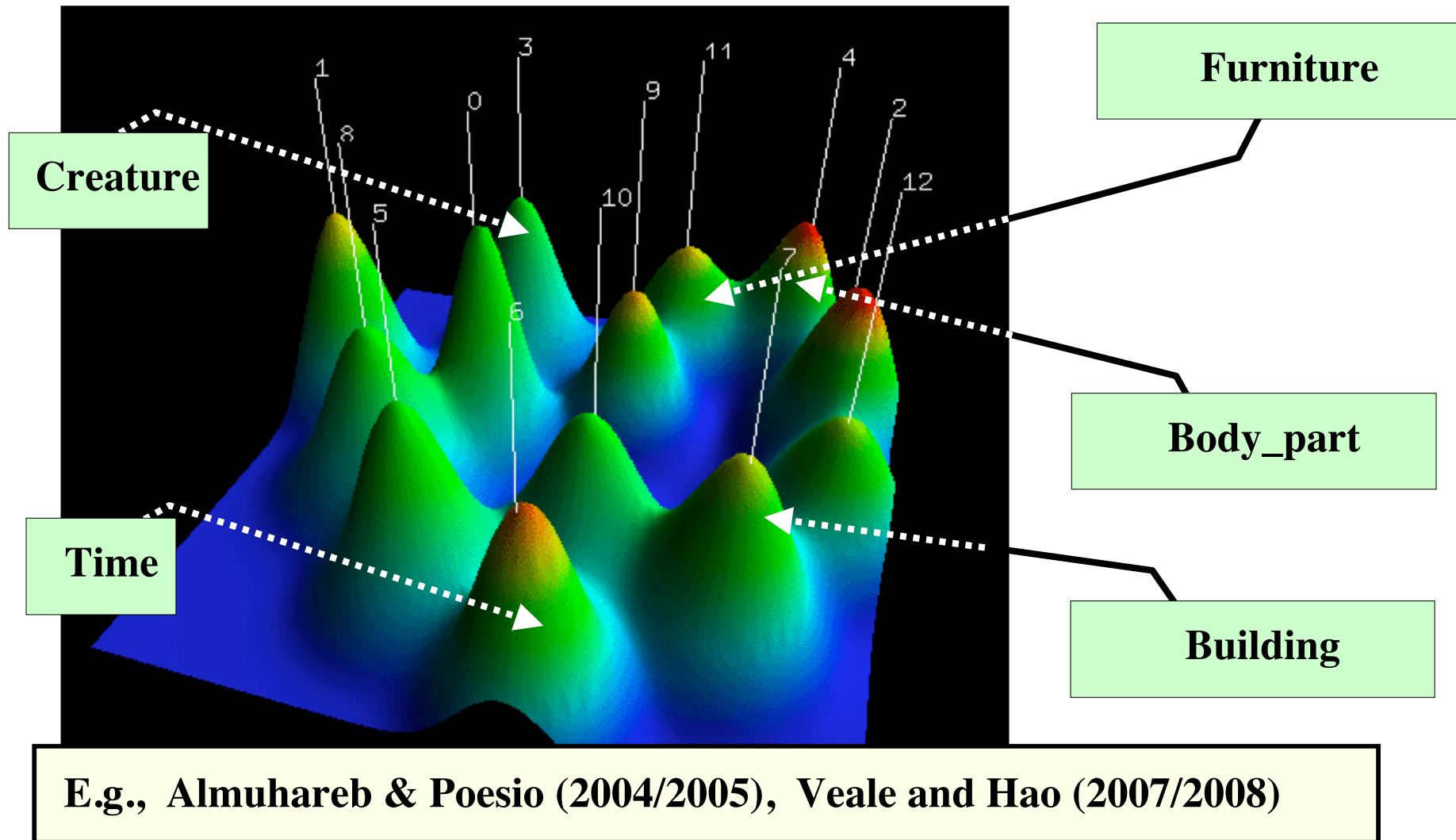
Analogy is a social connective: clique members have complementary roles

*Like* socializes with *Like*: Consider bands, teams, summits, parties, matches, etc.

*Like* socializes with *Like* → *Like* mingles with *Like* → *Like* collocates with *Like*

∴ Text Collocation → Category Co-Membership → Cliques as Categories

## Categories are tight-knit clusters of instances in a Feature Space



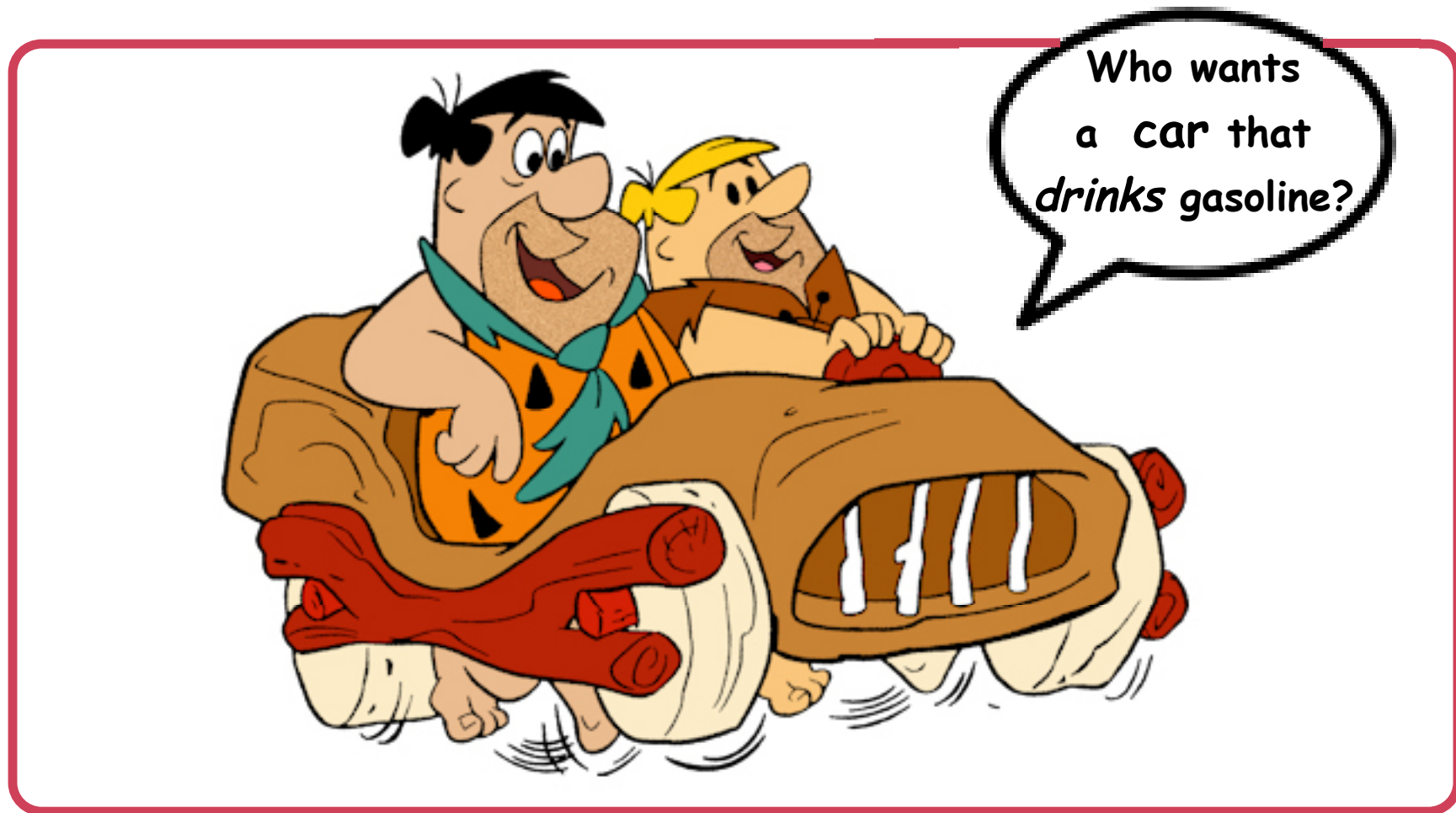
**Manipulating Categories: The View from Cognitive Science / Ling.**

# Metaphor

 **Lakoff**  
PRO RACING



## Modelling Metaphor: The State of the Art in Computer Science



**Fragile Systems / Tiny Coverage / Deep Knowledge Bottleneck / No Scalability**

## More Credible (& Useful) Goals: Lower Level, Broader Coverage



### A Modest Proposal

**Creative Information Mgmt.**  
*(texts, corpora, ngrams)*



**Creative Knowledge Mgmt.**  
*(categories, ontologies)*

*Creative information retrieval*  
*Metaphor finder / thesaurus*  
*Lexicalized Idea Exploration*

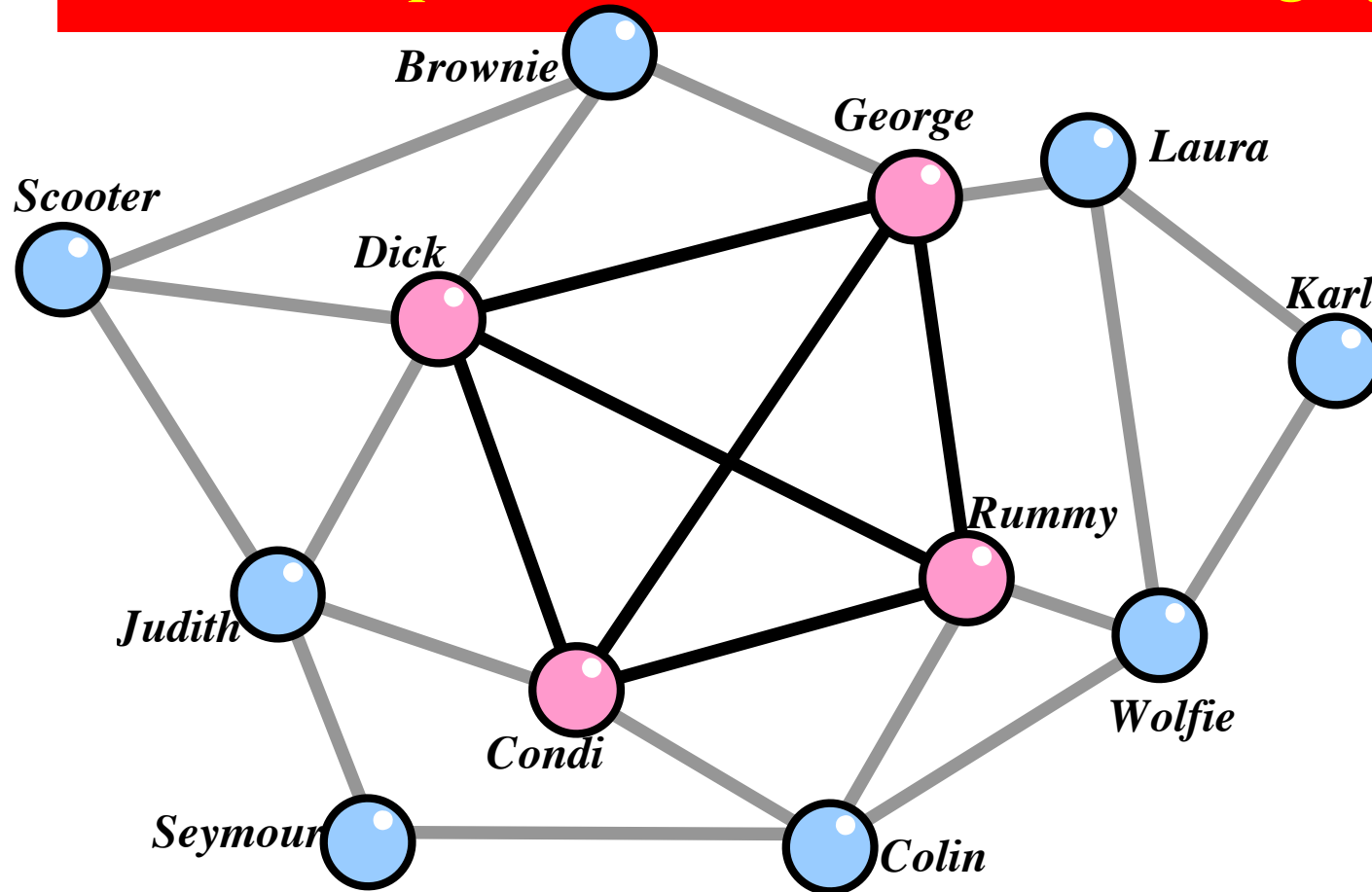
**Robust Systems / Broad Coverage / Plentiful Shallow Knowledge / Scalability**

## The Supermarket Principle: Group by Theme *and* Kind



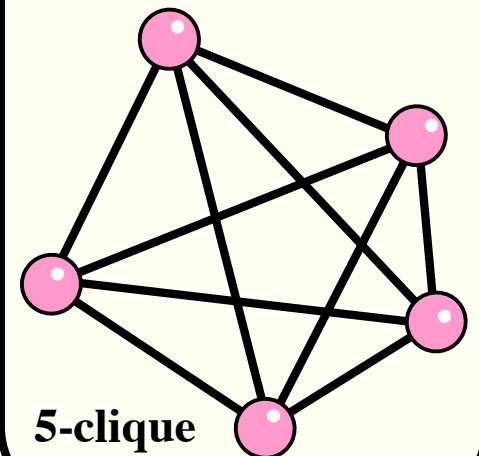
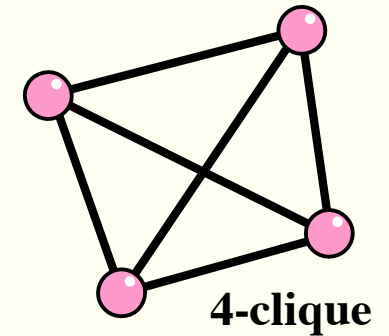
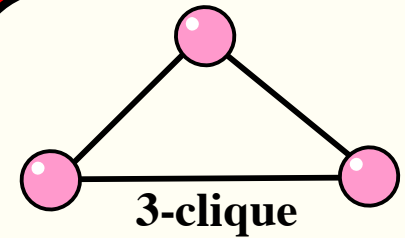
**Ontological Categories should reflect both semantic relatedness and similarity**

## Cliques: A Social-Model of Category Structure



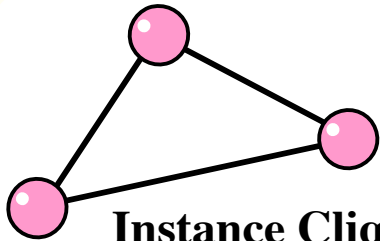
A Clique is a complete sub-graph.

A clique is Maximal iff *not* part of a larger clique.





## Linguistic Cliques: **Different Levels of** Conceptual Grouping

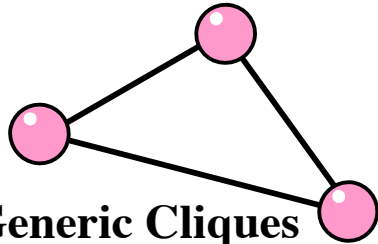


Instance Cliques

### Cliques of Proper-Name instances

Identified by capitalization patterns, Named-Entity Recognition

E.g., *Tiger Woods, Roger Federer and Thierry Henry*

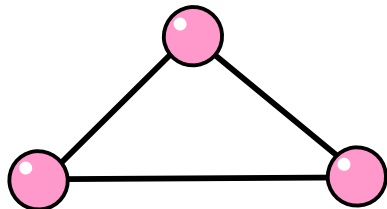


Generic Cliques

### Cliques of Generic Category References

Identified by patterns of coordination among *bare plurals*

E.g., *astronauts, cowboys and pirates*    *churches and priests*



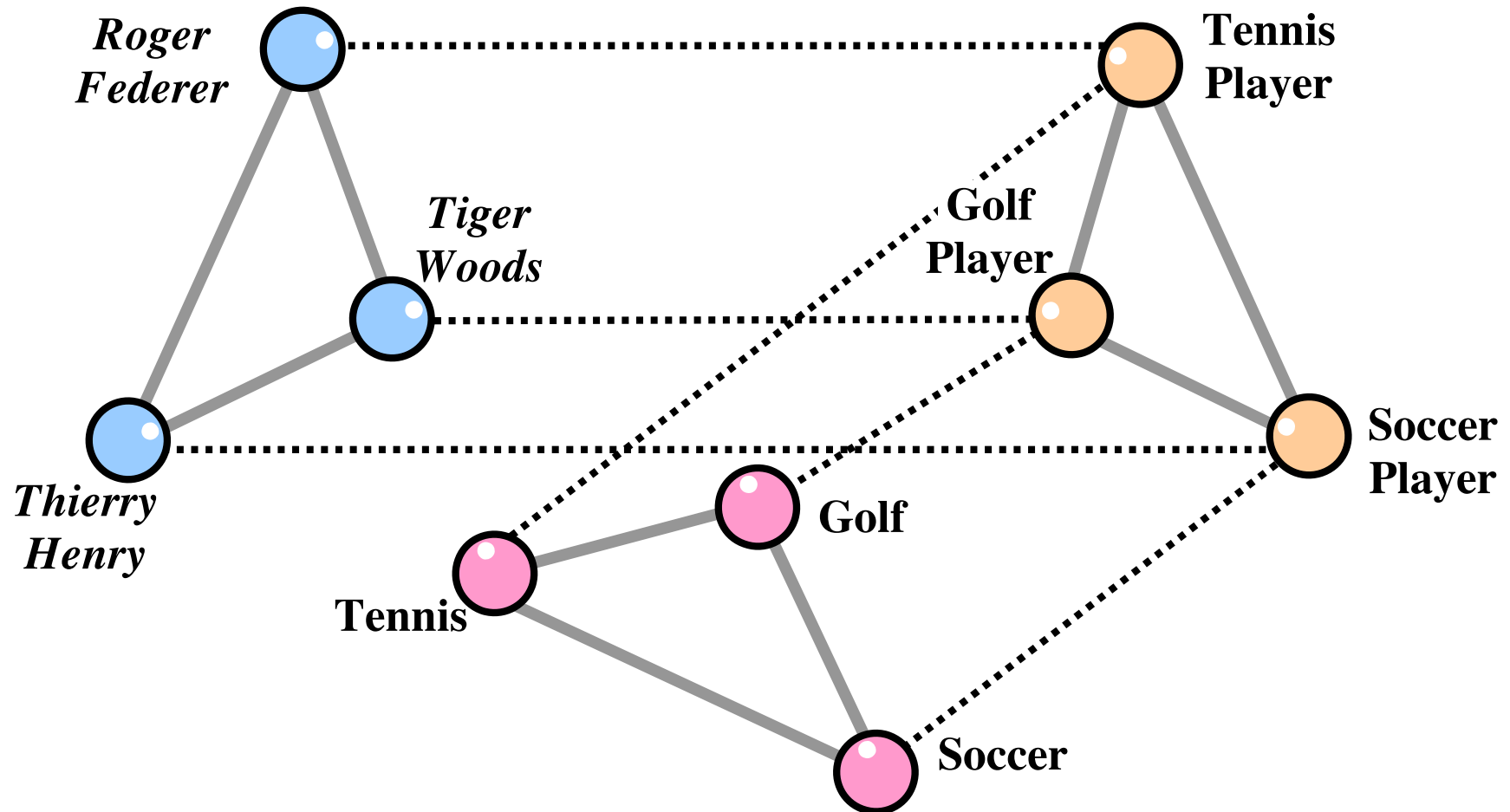
Category Cliques

### Cliques of Complex Category Names

Higher-level cliques formed from *generic and instance* cliques.

E.g., *Catholic church*    *and*    *Catholic priest*

## Analogy as a Social Connective: **Cliques imply Cliques**



Text Collocation → Instance cliques → Category cliques → Domain cliques

# Mining Collocations from Corpora: *Instances, Categories & Domains*

## 5-grams

*Roger Federer , Tiger Woods*  
Rafael Nadal and Roger Federer  
Roger Federer, Andy Roddick  
*Thierry Henry , Roger Federer*  
*Tiger Woods , Roger Federer*  
David Beckham, Thierry Henry  
Tom Cruise & David Beckham  
Tom Cruise and Katie Holmes  
Steven Spielberg, Tom Cruise  
Tom Hanks / Steven Spielberg  
Dan Brown and Tom Hanks  
*Tiger Woods vs. Thierry Henry*  
: : : : :

## 4-grams

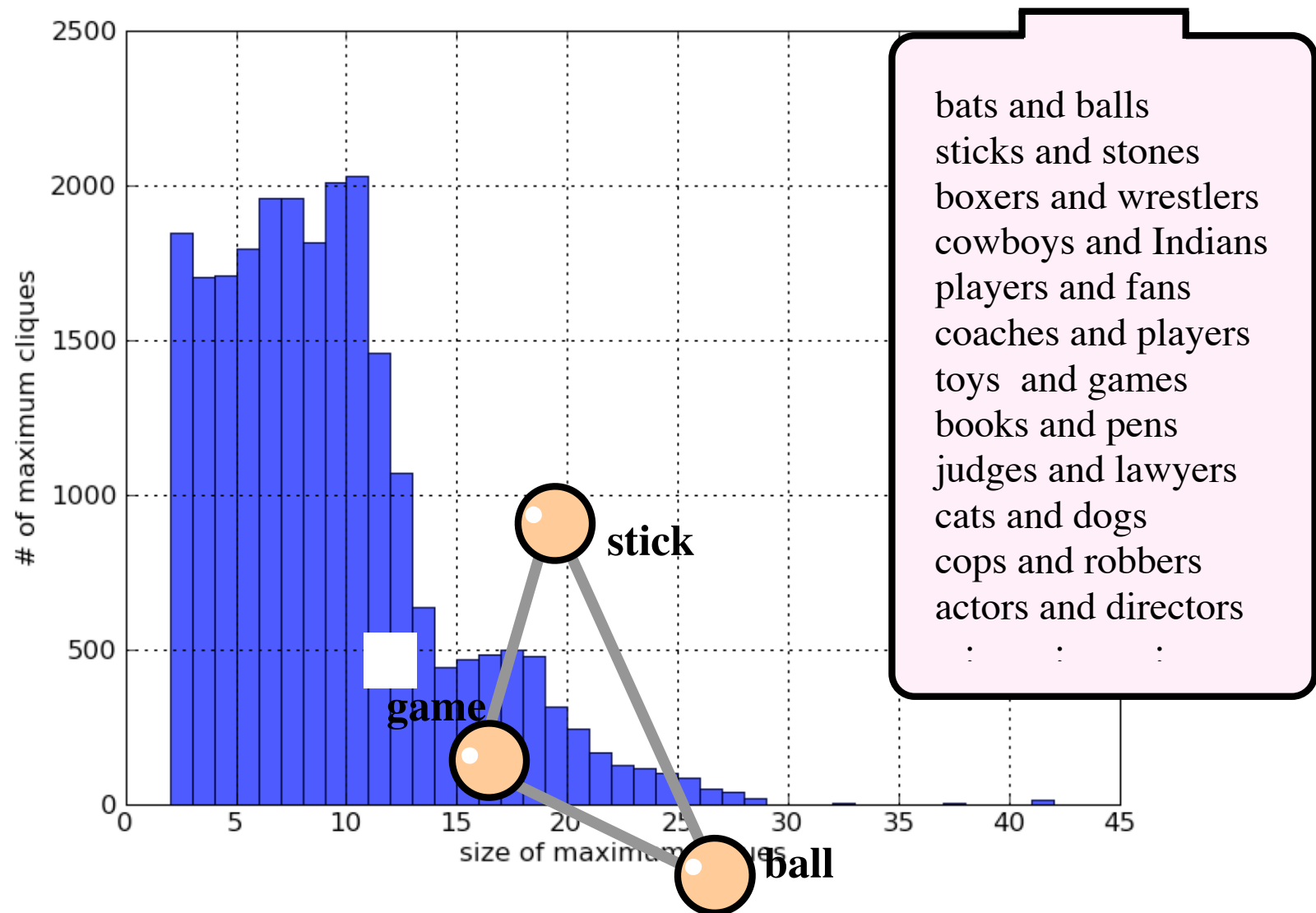
*tennis and golf players*  
tennis / squash players  
soccer and hockey moms  
polo and tennis teams  
squash and tennis courts  
soccer and rugby fields  
tennis and soccer fans  
*soccer and tennis players*  
polo and lacrosse teams  
*soccer vs. golf players*  
TV and movie stars  
radio and TV stars  
: : : :

## 3-grams

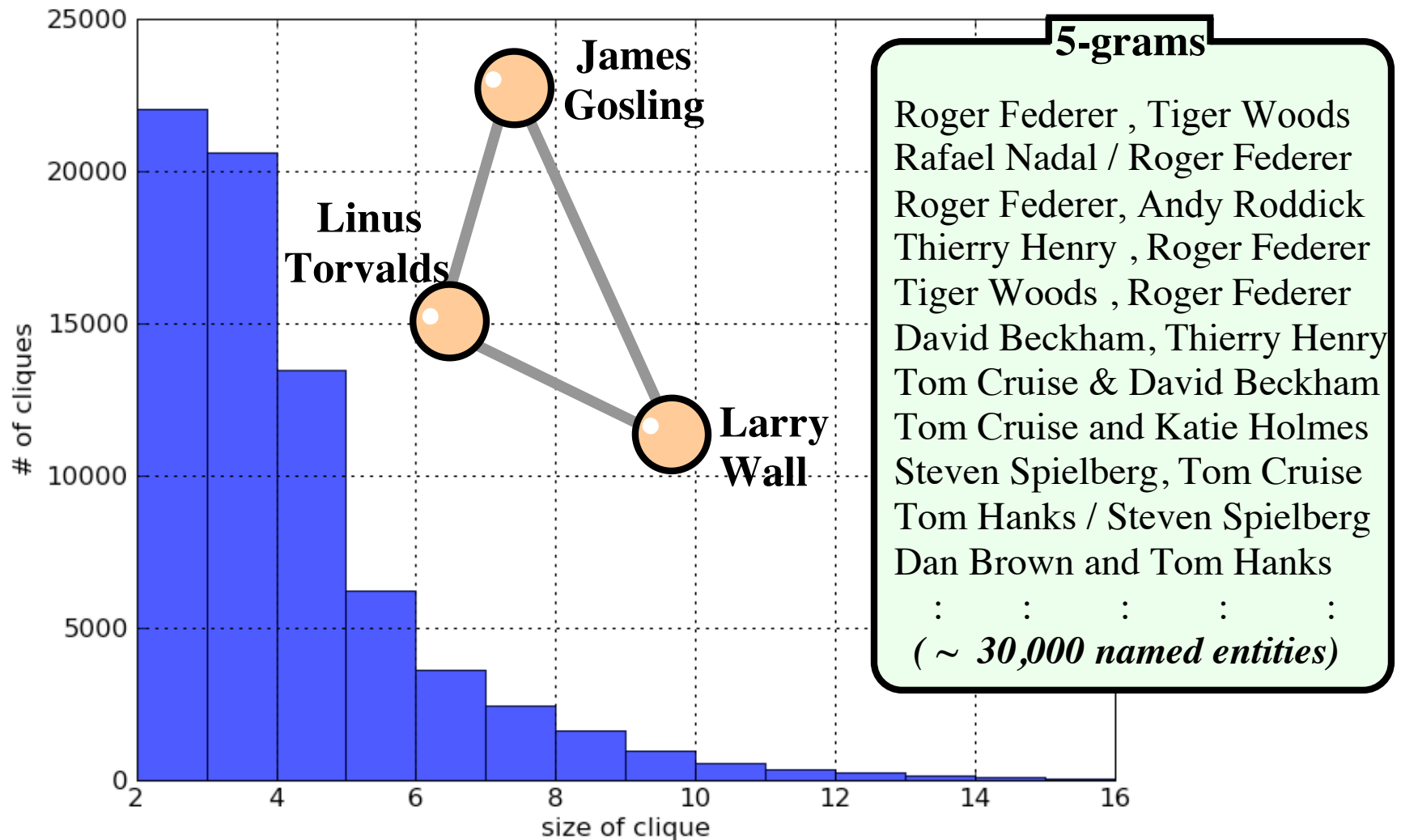
*tennis and golf*  
polo and tennis  
hockey and boxing  
boxing and karate  
players and fans  
coaches and players  
*golf vs. soccer*  
soccer and rugby  
*soccer versus tennis*  
Hollywood / Bollywood  
radio and TV  
actors and directors  
: :  


E.g., we use the Google N-Grams 1T Web Corpus ( $N \leq 5$ )

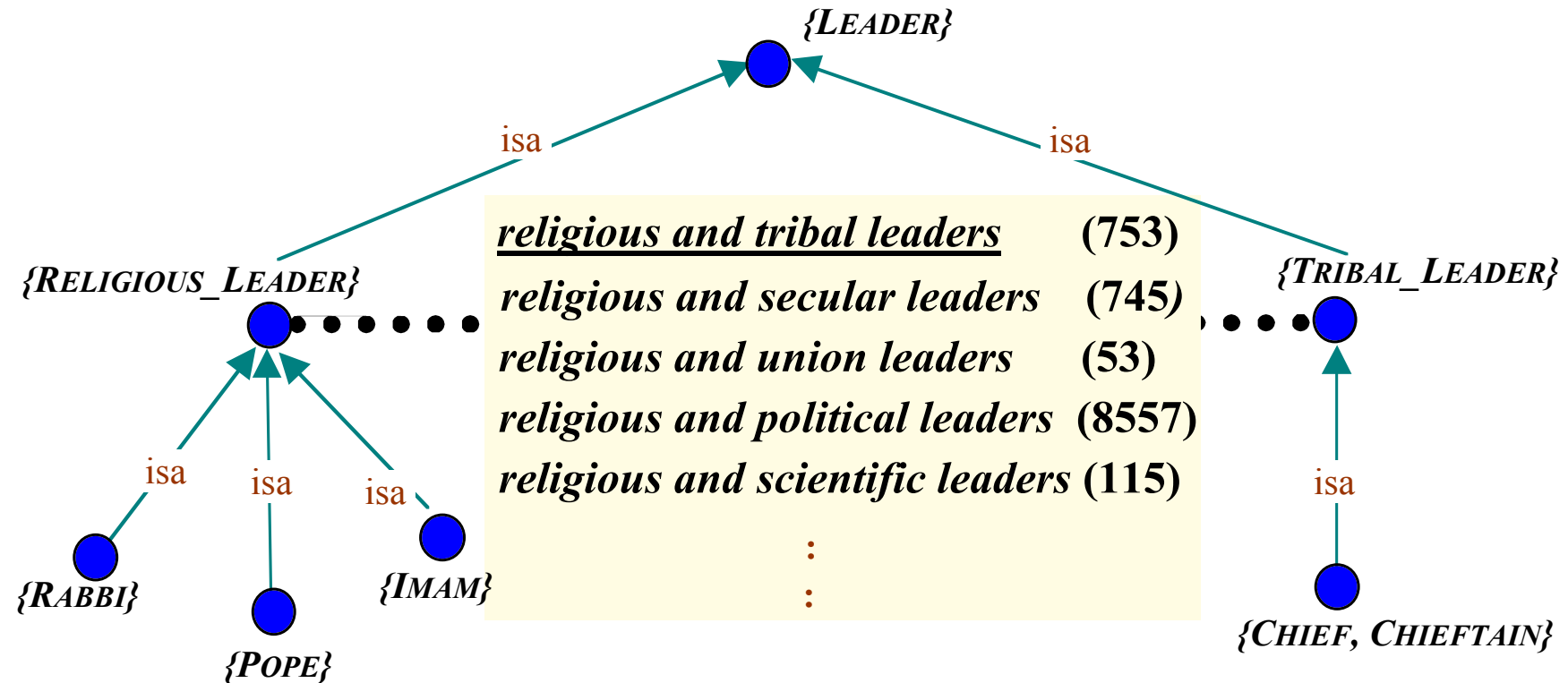
# Generic Clique Distribution in Corpus Data (coordinated plurals)



## Instance Clique Distribution in Corpus Data (proper individuals)

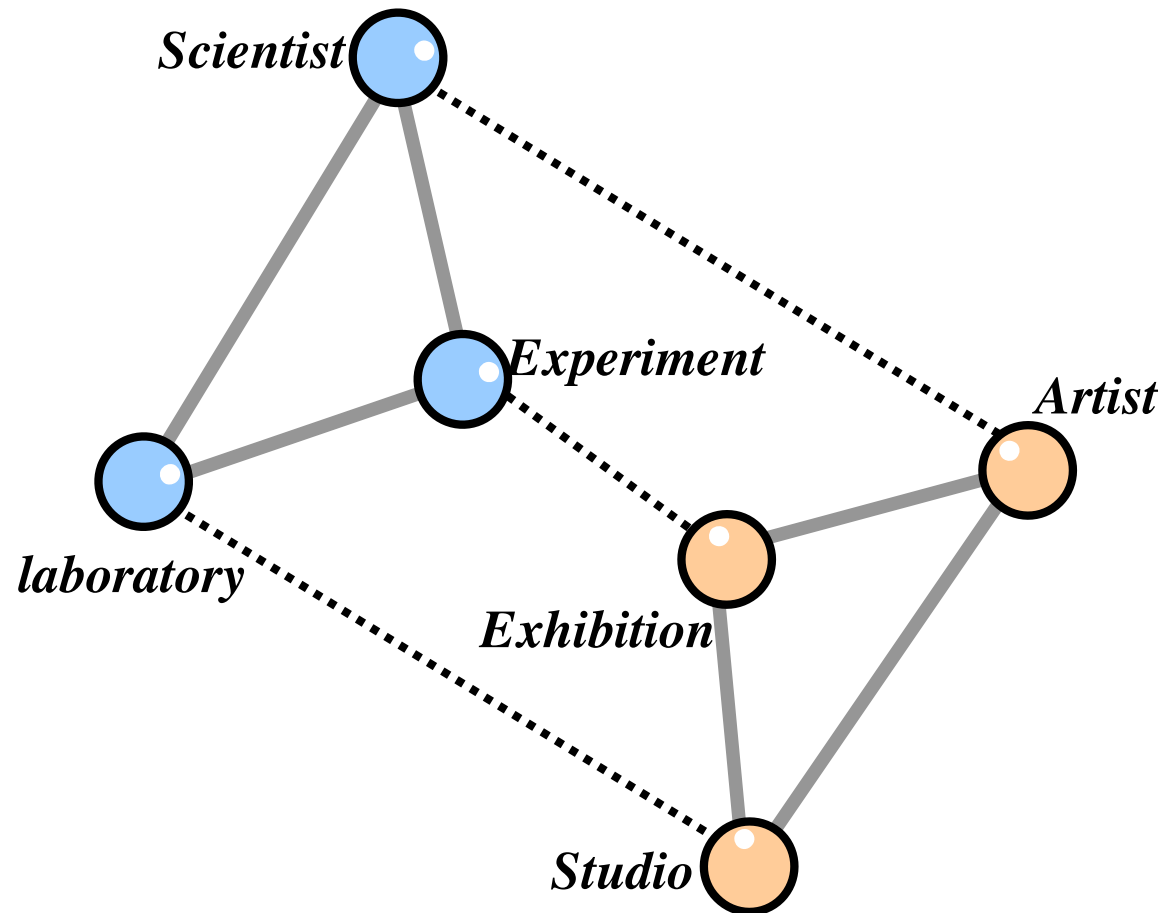


# Learning **Inter-Category Connections** from Corpus Cliques



Assume No Zeugma in compressed coordinations → metaphoric pathways

## Systematic Mapping between Cliques: **Cliques** on top of **Cliques**

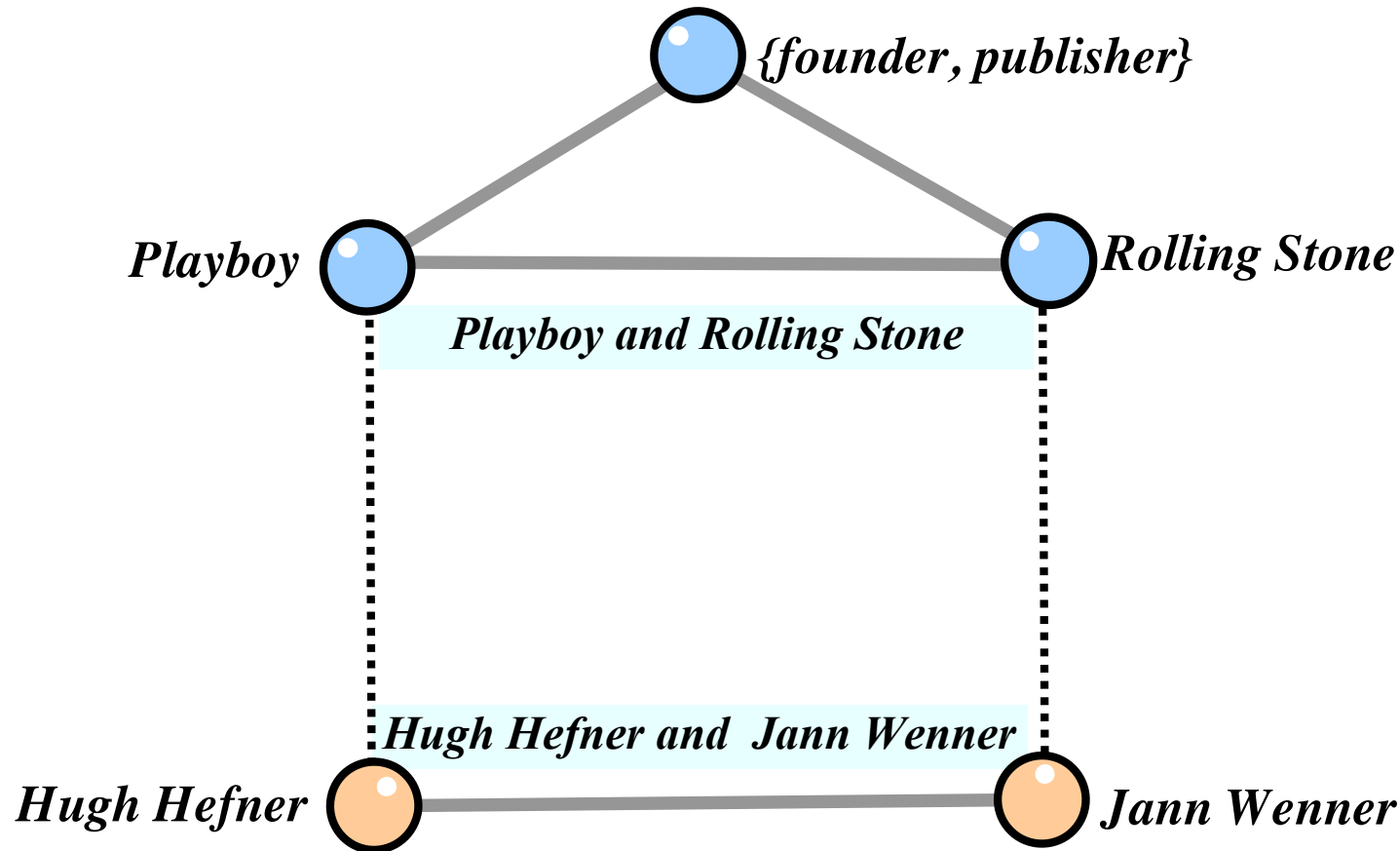


### 3-grams

*scientists and artists*  
*experiments & scientists*  
*studios and artists*  
*scientists and laboratories*  
*laboratories / experiments*  
*artists and exhibitions*  
*laboratories / studios*  
*studios and exhibitions*  
*exhibitions , experiments*  
*scientists and experts*  
*portfolios and artists*  
*chefs and kitchens*  
:  
:  
:

Text Collocation → 2-clique *counterparts* & 3-clique *core representations*

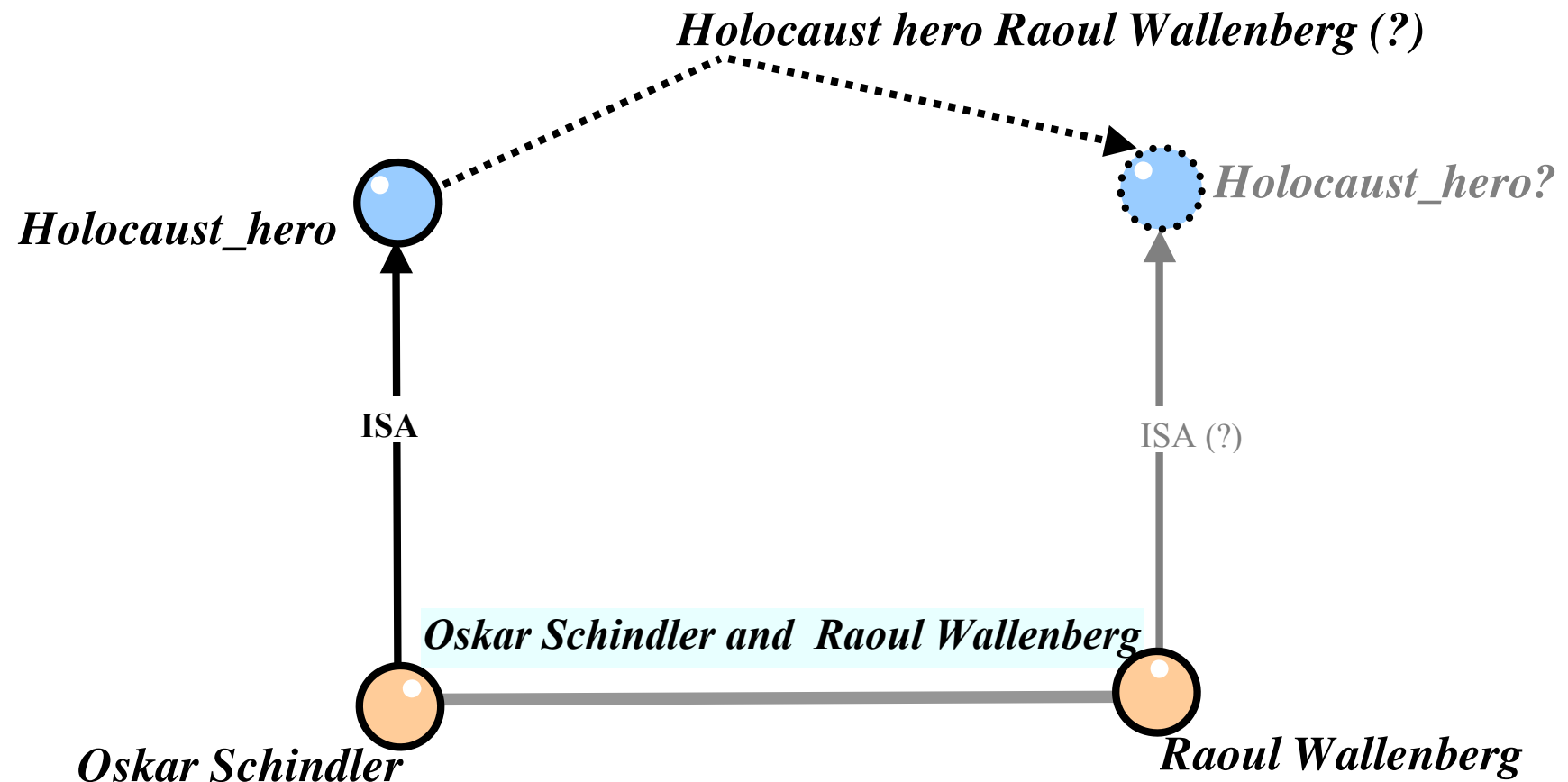
## Analogy as a Social Connective (II): Mapping Like-to-Like



Textual Collocation of Instances → Analogy between collocated Categories

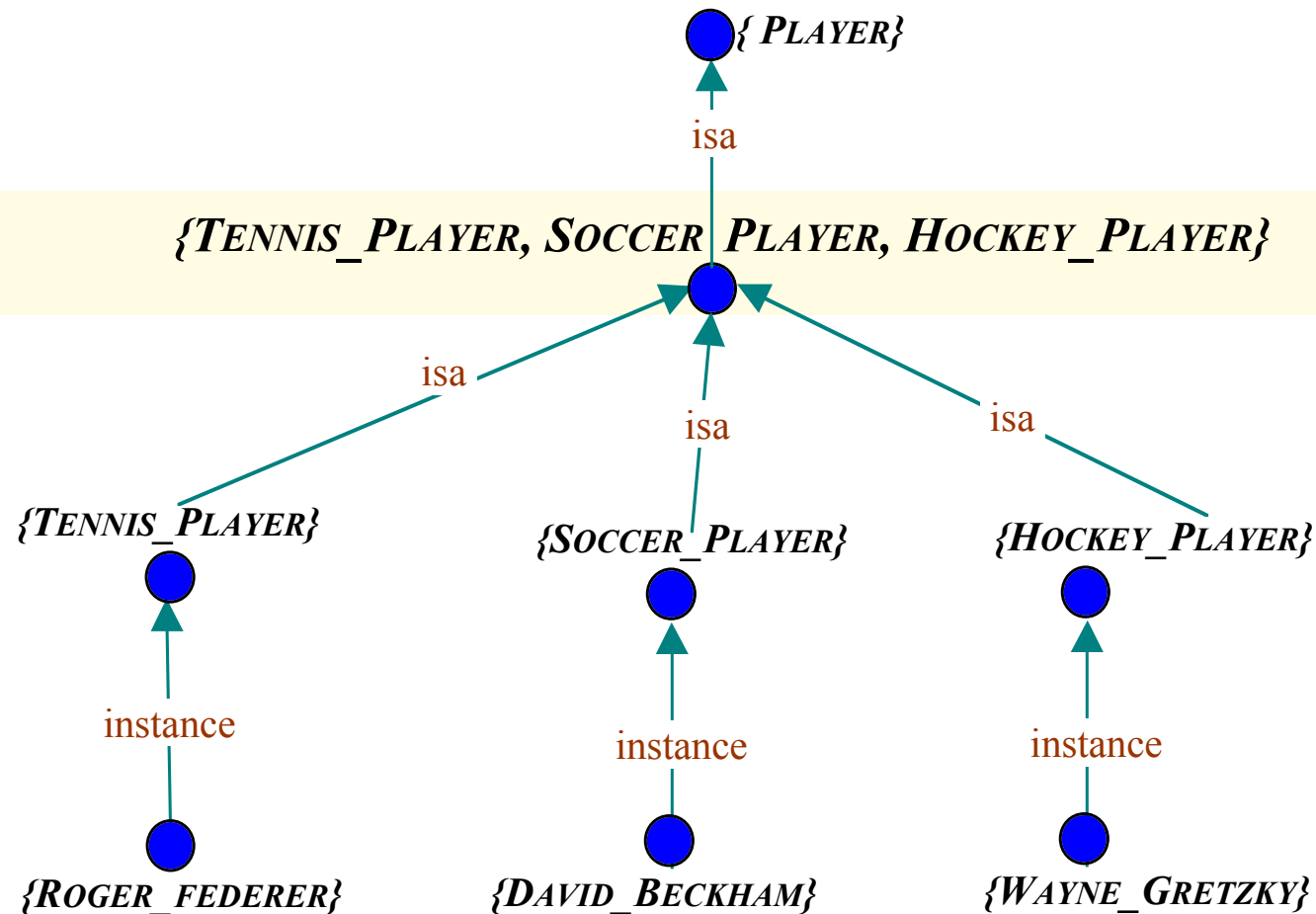


## Analogy as a Social Connective (III): **Analogical Bootstrapping**



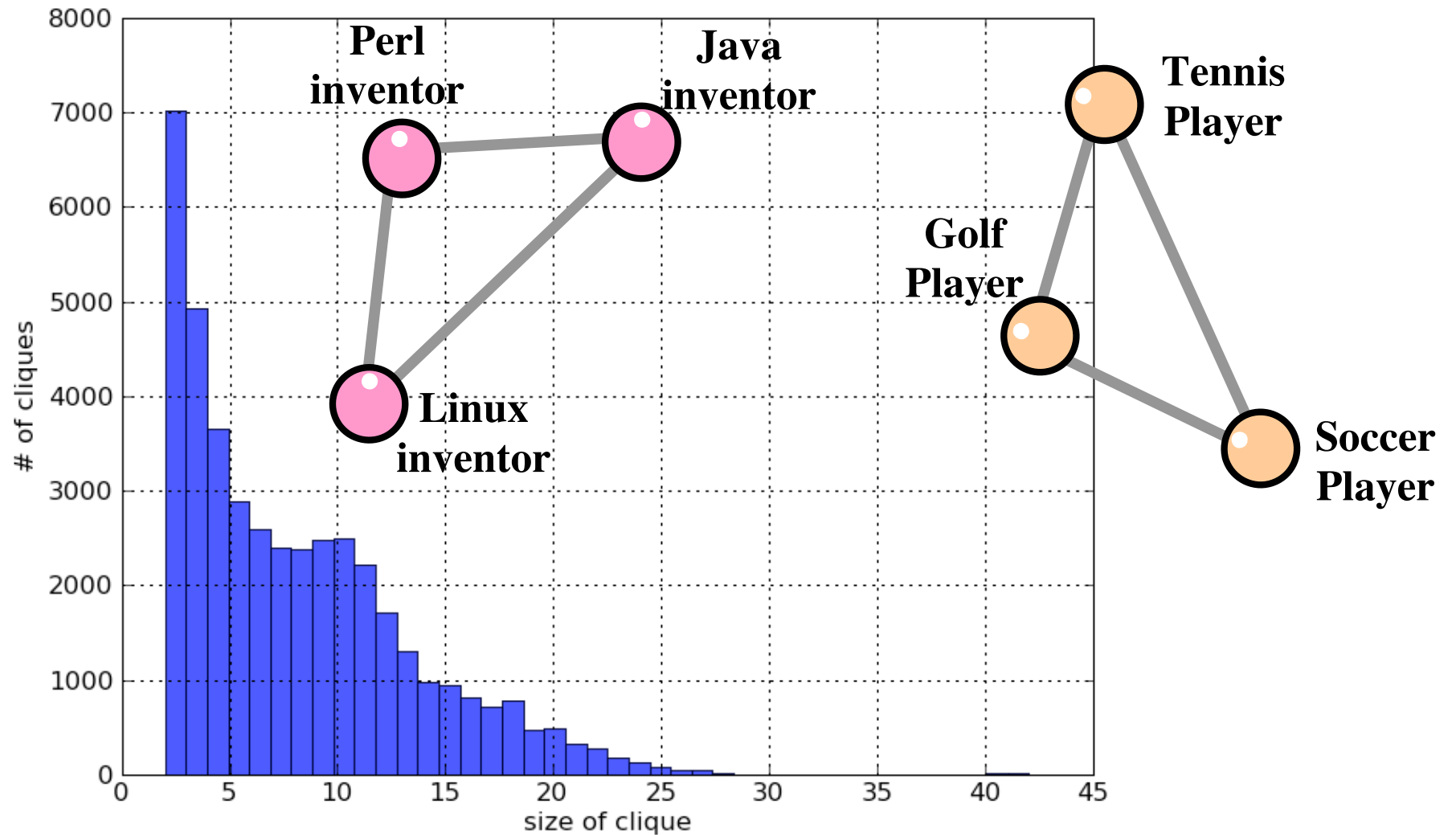
Collocation of Instances → Shared Category (?) → Bootstrapping Hypothesis

## Cliques As **Categories-of-Categories** for Additional Structure

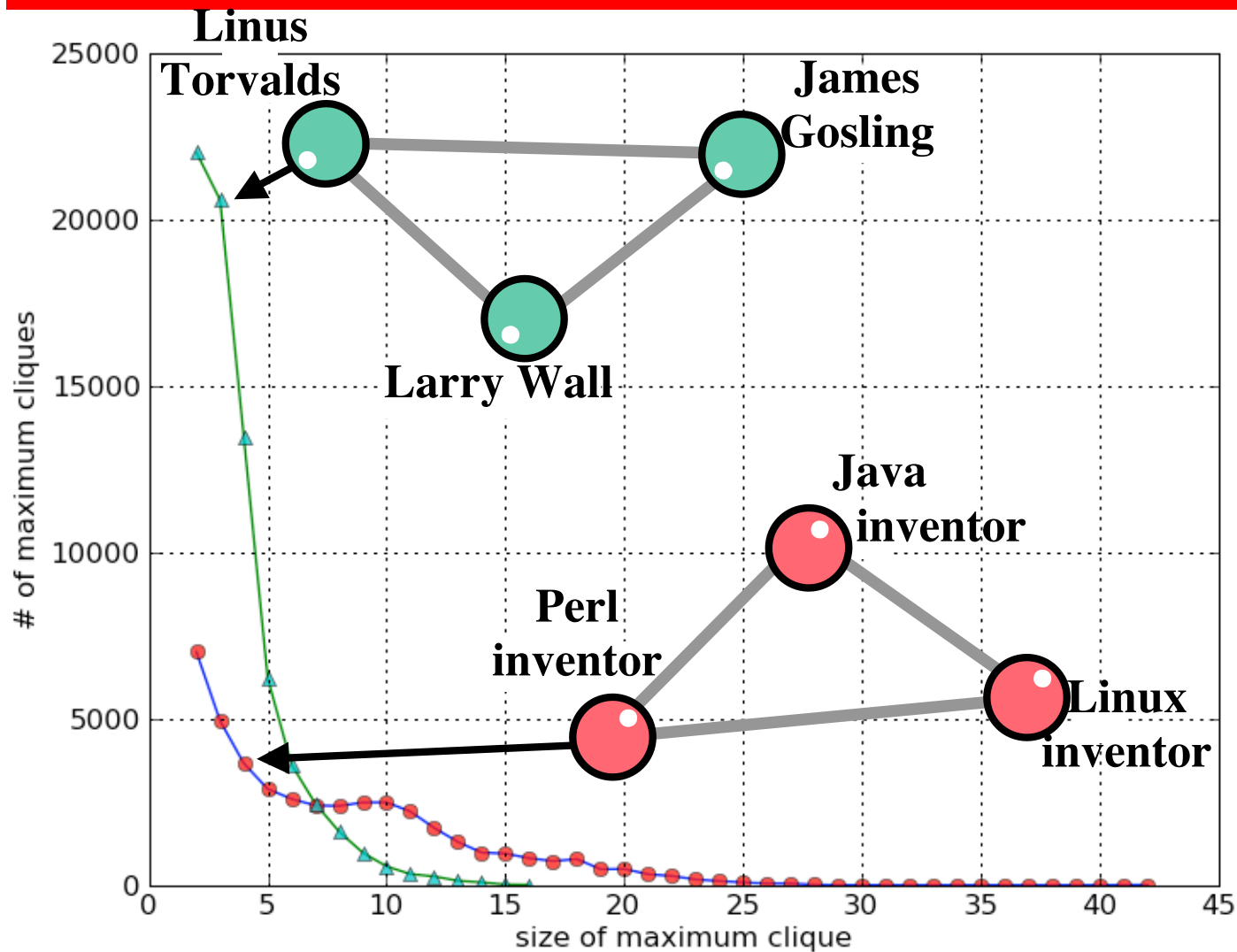


Insert Cliques of Categories as New Upper-Ontology Categories in WordNet

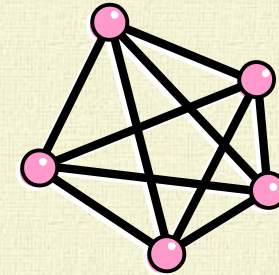
## Clique Distribution at Category Level (analogical mappings)



## Comparing Clique Distributions (instance- and analogical- cliques)



Afflatus.UCD.ie/dorian



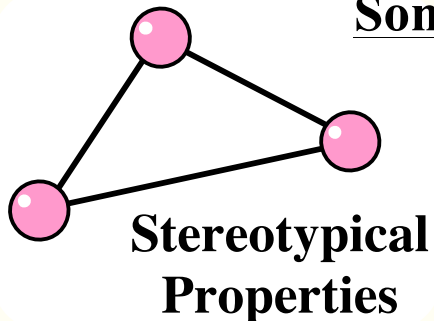
## Portraits by Dorian

[ [JACK\\_Kirby](#) [Jack\\_Kirby](#) ...15 omitted... [Venus\\_Williams](#) [Martina\\_Hingis](#) [Andre\\_Agassi](#) [Roger\\_Federer](#) ] [full history>>>](#)

### Instance: Roger\_Federer

Friends	Categories	
<input type="checkbox"/> <a href="#">Martina Hingis</a>	<input type="checkbox"/> <a href="#">US Open champion</a>	+ <input type="checkbox"/> <a href="#">Marcos Baghdatis</a>
<input type="checkbox"/> <a href="#">Kim Clijsters</a>	<input type="checkbox"/> <a href="#">Wimbledon champion</a>	- <input type="checkbox"/> <a href="#">Michael Jordan</a>
<input type="checkbox"/> <a href="#">Lleyton Hewitt</a>	<input type="checkbox"/> <a href="#">French Open champion</a>	• <a href="#">Basketball player</a> +
<input type="checkbox"/> <a href="#">Andy Roddick</a>	<input type="checkbox"/> <a href="#">Australian Open champion</a>	• <a href="#">basketball player</a> +
<input type="checkbox"/> <a href="#">Andre Agassi</a>	<input type="checkbox"/> <a href="#">tennis player</a>	• <a href="#">basketball star</a> +
<input type="checkbox"/> <a href="#">Rafael Nadal</a>	<input type="checkbox"/> <a href="#">Open champion</a>	• <a href="#">basketball superstar</a> -
<input type="checkbox"/> <a href="#">Maria Sharapova</a>	<input type="checkbox"/> <a href="#">tennis star</a>	◦ <a href="#">tennis player</a>
<input type="checkbox"/> <a href="#">Venus Williams</a>	<input type="checkbox"/> <a href="#">Laureus World Sports Awards winner</a>	◦ <a href="#">tennis star</a>
<input type="checkbox"/> <a href="#">Lindsay Davenport</a>	<input type="checkbox"/> <a href="#">Tennis player</a>	+ <input type="checkbox"/> <a href="#">Tim Henman</a>
<input type="checkbox"/> <a href="#">Yelena Isinbayeva</a>	<input type="checkbox"/> <a href="#">Swiss tennis player</a>	+ <input type="checkbox"/> <a href="#">Andy Murray</a>
<input type="checkbox"/> <a href="#">Carlos Moya</a>	<input type="checkbox"/> <a href="#">tennis champion</a>	+ <input type="checkbox"/> <a href="#">John Daly</a>
		+ <input type="checkbox"/> <a href="#">Kobe Bryant</a>
		+ <input type="checkbox"/> <a href="#">Max Mirnyi</a>

## Salient Properties: **Clique Members often share** defining properties



Some members of a clique have strong stereotypical associations

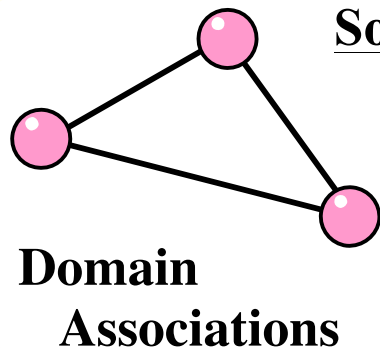
These properties are then sensibly projected onto co-members

E.g., *Cowboys, Astronauts and Pirates*

reckless

brave

bold



Some members belong to different, but parallel, domains

The domains are then sensibly projected onto co-members

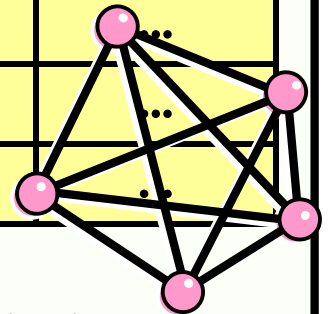
E.g., *Astronauts and space scientists*

E.g., *Astronauts and NASA officials*

## Category **Properties** also socialize in **Clique** Structures

Adjacency matrix of *mutually-reinforcing* properties acquired from WWW:

	hot	spicy	humid	fiery	dry	sultry	...
hot	---	35	39	6	34	11	...
spicy	75	---	0	15	1	1	...
humid	18	0	---	0	1	0	...
fiery	6	0	0	---	0	0	...
dry	6	0	0	0	---	0	...
sultry	11	1	0	2	0	---	...
...	...	...	...	...	...	...	...

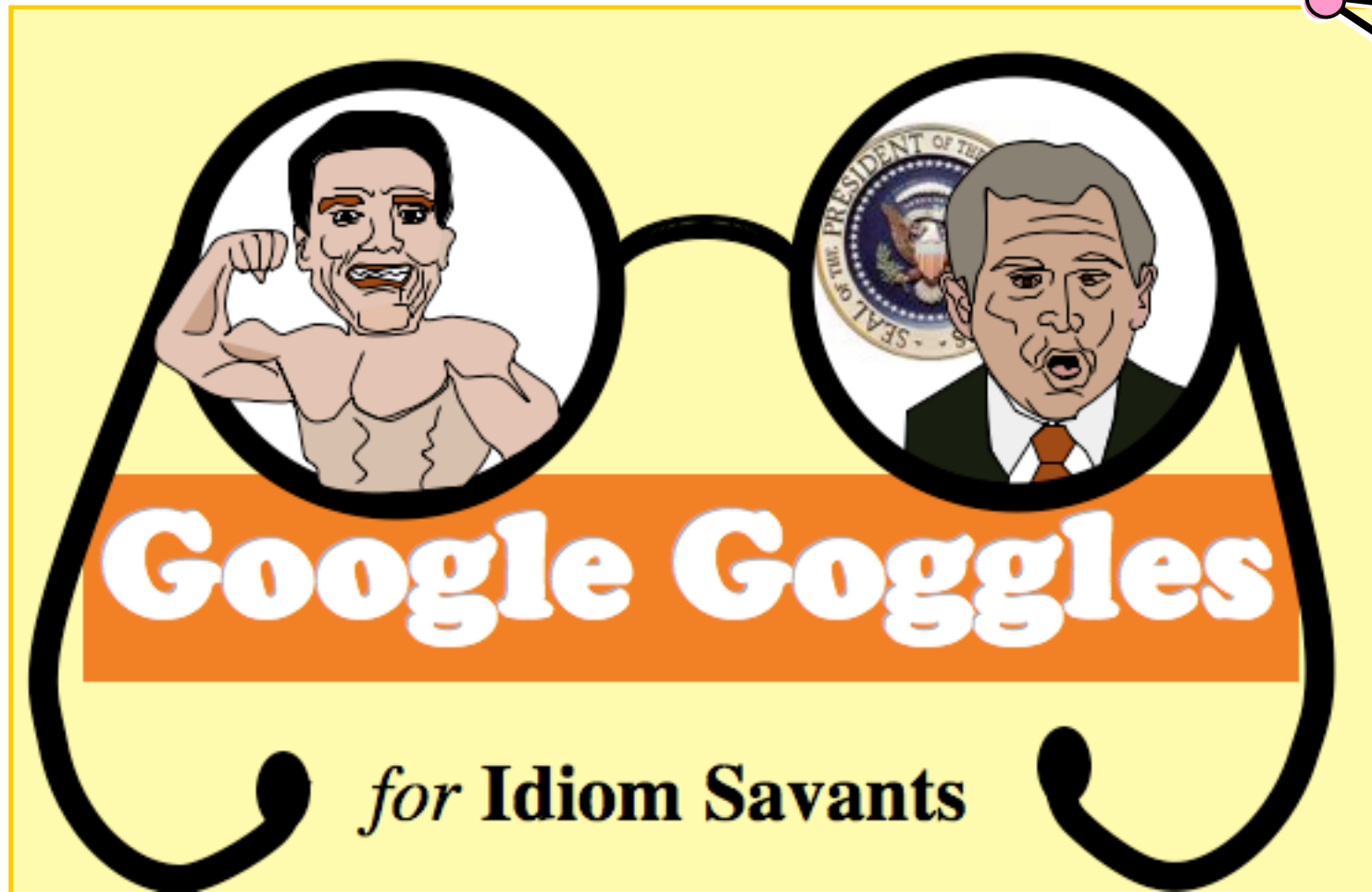
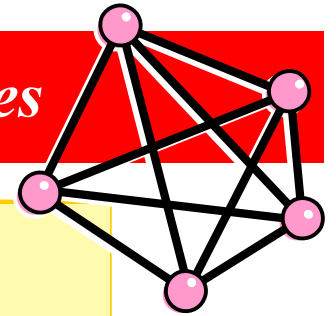


Use the Google query

**“as \* and \* as”**

to acquire associations

## Creative Information Retrieval: *Cliques as Categories*



View the Google 1T corpus of ngrams as a *Lexicalized Idea Space* to Explore





**45 2-grams Retrieved:**

Muslim	?rabbi
--------	--------













☒ Order by Frequency/Similarity   ☐ Case sensitive

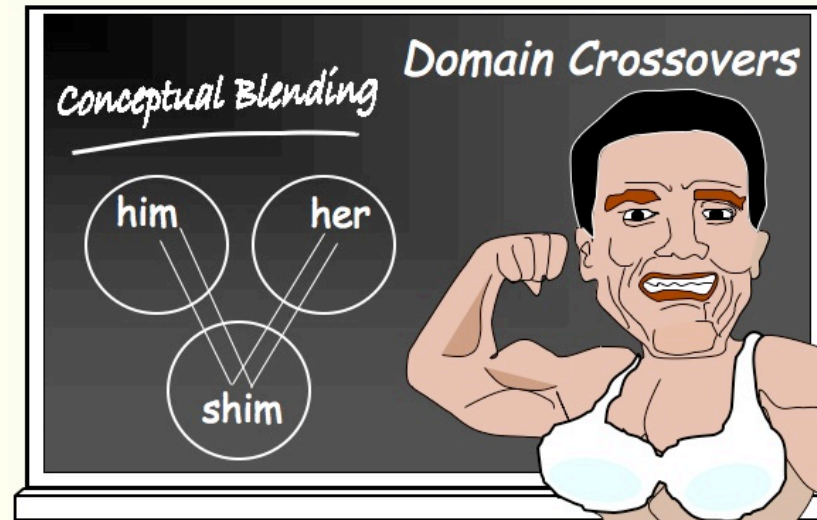
Submit

Export

Export

CrossOvers

Gram 1	Gram 2	Freq	Sim
Muslim	priest	979	87  
Muslim	leader	38403	76  
Muslim	minister	761	74  
Muslim	preacher	6140	71  
Muslim	imam	1908	70  
Muslim	cleric	76212	70  



## Type A Keyword

astronaut

☐ Case sensitive

Submit

### 1 Stereotypes

adventurous astronaut

### 8 Expected Forms

NASA astronaut

Russian astronaut

space astronaut

Soviet astronaut

mission astronaut

flight astronaut

rocket astronaut

senior astronaut

### 0 Common Uses

## 40 2-grams Retrieved:

space	?astronaut
-------	------------































☒ Order by Frequency/Similarity ☐ Case sensitive

Submit

Export

Export

CrossOvers















Gram 1	Gram 2	Freq	Sim
space	astronaut	445	100  
space	alien	18514	81  
space	satellite	7050	70  
space	researcher	243	67  
space	worker	208	66  
space	technician	62	66  
space	specialist	307	66  
space	physicist	1024	66  
space	official	2315	66  
space	employee	140	66  
space	diver	69	66  
space	cowboy	9988	66  
space	artist	1437	66  
space	angel	181	66  
space	scientist	5174	65  

## 297 2-grams Retrieved:


☒ Order by Frequency/Similarity ☐ Case sensitive





Gram 1	Gram 2	Freq	Sim
space	scientist	5174	100  
space	physicist	1024	94  
space	biologist	49	93  
space	researcher	243	91  
space	guide	457	63  
space	geek	982	63  
space	fan	1127	63  



## 2 4-grams Retrieved:

kitchen	-	*ed	?kitchen
---------	---	-----	----------

☒ Order by Frequency/Similarity ☐ Case sensitive

Submit

Export	Export	Export	Export		
Gram 1	Gram 2	Gram 3	Gram 4	Freq	Sim
kitchen	-	Fitted	kitchen	370	100 →
kitchen	-	tested	recipe	104	1 →

## 28 4-grams Retrieved:

?kitchen	-	tested	?recipe
----------	---	--------	---------

☒ Order by Frequency/Similarity ☐ Case sensitive
















Submit

Export

Export

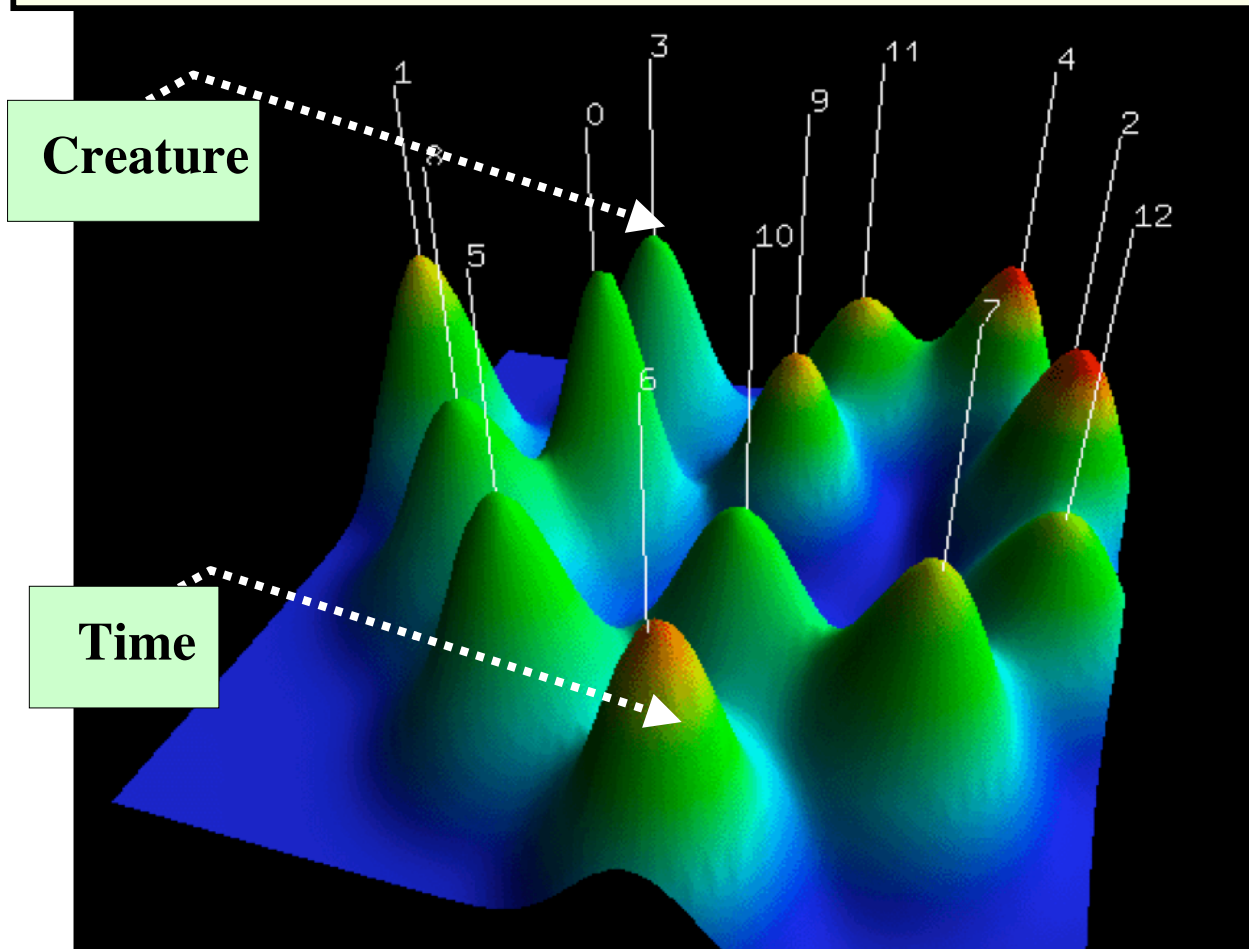
Export

Export

Gram 1	Gram 2	Gram 3	Gram 4	Freq	Sim
kitchen	-	tested	recipe	104	100 
classroom	-	tested	lesson	212	52 
classroom	-	tested	program	88	50 
classroom	-	tested	advice	44	50 
classroom	-	tested	code	51	34 
market	-	tested	application	63	26 
field	-	tested	solution	5824	26 
field	-	tested	lesson	84	26 
field	-	tested	program	140	25 
field	-	tested	collection	72	25 
field	-	tested	advice	357	25 
laboratory	-	tested	formula	76	22 
kitchen	-	tested	holiday	59	21 
classroom	-	tested	model	56	20 
market	-	tested	product	62	15 

# Feature Space and Categorization Via Clustering: **Revisited**

13-way clustering of 214 nouns, compared to WordNet



Almuhareb & Poesio (2004)

*Weak text-derived features*

~ 60,000 features for 214 nouns

Result: 0.855 cluster purity

Veale & Hao (2007 / 2008)

*Strong simile-derived features*

~ 7,300 features for 214 nouns

Result: 0.902 cluster purity

Veale & Li (2009)

*Generic Clique-derived features*

~ 8,300 features for 214 nouns

Result: 0.934 cluster purity





**MOEDfinger**

[Log in](#)

[Create an account](#)

Europe -anxiety

[Search This!](#)

[Show word cloud](#)

Show 792 results

Sort by

[date](#)

[density](#)



[Comments: Nearly 40 percent of Europeans suffer mental illness](#) - Reuters - 2011-09-04

Mood words: [depression](#) [turmoil](#) [trouble](#) [doubt](#) [stress](#) [panic](#) [disappointment](#) [illness](#)

Social Stream (What's this?) Wish they compared **Europe** to the U.S., Asia and other countries. ~ I have no doubts that 40% of Europeans are crazy, because it is more like 80% of them are crazy ...



[Greece: Depression and Suicide Rates Soar as Economy Dives - Yahoo! News](#) - Yahoo News - 2011-06-27

Mood words: [suicide](#) [depression](#) [panic](#) [crisis](#) [anxiety](#)

the rest of **Europe** may be tormented by the thought of having to cough up ever more money to bail out Athens, the once carefree Greeks are getting more depressed by the day. Psychiatrists say ...



[Andrew Wigley: Authority & Authoritarianism: Les révoltés du «no future»](#) - Huffington Post - 0000-00-00

Mood words: [hardship](#) [chaos](#) [depression](#) [vulnerability](#) [agitation](#) [violence](#) [fear](#) [debilitating](#) [unrest](#) [panic...](#)

50 years ago this week construction began on the Berlin Wall. The wall came to symbolise the rigid divisions in ideology between Western and Eastern **Eui** and the brutal, uncompromising ...



[Eurozone Crisis: Merkel And Sarkozy Reaffirm Commitment To Greece](#) - Huffington Post - 2011-09-14

Mood words: [fight](#) [danger](#) [chaos](#) [turmoil](#) [crisis](#) [debt](#) [fear](#) [scepticism](#) [doubt](#) [panic...](#)

was heading for bankruptcy caused panic in the markets on Wednesday, with Jacek Rostowski, the Prime Minister of Poland, suggesting that the entire European project was in jeopardy. **Europe** is in danger," he ...



[Europe's MASSIVE Austerity Measures: The Countries Making The Biggest Cuts \(PHOTOS\)](#) - Huffington Post - 2010-06-29

Mood words: [risk](#) [debt](#) [opposition](#) [depression](#)

Huffington Post | Nathaniel Cahners Hindman First Posted: 06-29-10 03:53 PM | Updated: 06-29-10 07:28 PM To reduce their budget deficits, nations





**MOODfinger**

[Log in](#)

[Create an account](#)

Korea -anxiety

Search This!

## Show word cloud

Show 800 results

Sort by

date

density



[Koreans made online suicide pact | World news | The Guardian](#) - Guardian - 2004-03-24

Mood words: [suicide](#) [death](#) [depression](#) [conflict](#)

The bodies of the three men and two women were found by a hotel employee on Monday night, according to the **Korea** Times. Police believe the five, age between 19 and 29, died after taking a cocktail ...



[North Korean leader Kim Jong-il travels to Russia to meet Medvedev - Telegraph](#) - Telegraph - 2011-08-20

Mood words: [depression](#)

of the long-stalled six-party talks on North Korea's nuclear programme, though China now holds more influence with North **Korea** than Russia does. Kim made his last public visit to Russia in 2002, when ...



[Yeonpyeong Island: A history - Telegraph](#) - Telegraph - 2010-11-23

Mood words: [conflict](#) [war](#) [violence](#) [attack](#) [tension](#) [fear](#)

9:18PM GMT 23 Nov 2010 The archipelago, which North **Korea** laid claim to in the 1970s, lies in the Yellow Sea near the Northern Limit Line – the sea border between the two countries. Around 1,600 ...



[Rival Koreas gear up for talks this week to ease tension](#) - Reuters - 2011-02-06

Mood words: [conflict](#) [crisis](#) [war](#) [risk](#) [tension](#)

the crisis and restart international talks over the North's nuclear program. The two **Koreas** are still technically at war because an armistice not a treaty ended the 1950-53 Korean War. The meeting ...



[North Korea Blamed For Cyberattack On U.S., South Korea: Report](#) - Huffington Post - 2011-07-05

Mood words: [threat](#) [attack](#) [war](#) [denial](#) [problem](#)

LOLITA C. BALDOR 07/ 5/11 10:24 AM ET WASHINGTON -- North **Korea** or its sympathizers were responsible for the cyberattack against South Korean

[Log in](#)[Create an account](#)

### Hide word cloud

hesitation hysteria fret hubris hardship **fight** withdrawal antagonism **threat** upheaval peril indignation rage unhappiness unease doom hatred impatience urgency nagging aggression danger abuse burden resistance crushing suicide intolerance frustration confusion chaos disturbance discrimination death negativity apprehension insecurity blood depression hunger inadequacy hassle struggle crippling silence instability overwhelming conflict vulnerability alienation dislike sweat turmoil shame resignation harm strain annoyance deprivation exhaustion rejection nervousness pain

ignorance oppression paranoia trouble irritation outrage **crisis** agitation trial disquiet pity violence embarrassment **debt** paralysis anxiety indecision prejudice seizure dissatisfaction disillusionment injury attack indifference horror cancer avoidance **fear** greed hate ego envy remorse terror disruption difficulty fever despondency fatigue dependency grief dismay suspicion debilitating lethargy pessimism unproductive gloom scepticism guilt darkness misery torture unrest impotence doubt opposition stress friction displeasure criticism humiliation weakness **failure** disorder panic anger hostility inconvenience cynicism **loss** caution sadness disbelief protest compulsion desperation disease aversion discord **war** agony resentment **risk** disappointment timidity consternation inability poverty isolation nausea illness repression disability addiction uncertainty distress mistrust denial despair reluctance trauma loneliness malaise skepticism harassment regret **problem** illusion dependence misunderstanding shock noise tears distrust alarm headache weariness strife worry tension

→ [Eurozone Crisis: Merkel And Sarkozy Reaffirm Commitment To Greece](#) - Huffington Post - 2011-09-14

Mood words: [fight danger chaos turmoil crisis debt fear scepticism doubt panic...]

was heading for bankruptcy caused panic in the markets on Wednesday, with Jacek Rostowski, the Prime Minister of Poland, suggesting that the entire European project was in jeopardy. **Europe** is in danger," he ...

→ [Europe's MASSIVE Austerity Measures: The Countries Making The Biggest Cuts \(PHOTOS\)](#) - Huffington Post - 2010-06-29

Mood words: [risk debt opposition depression]

[Log in](#)[Create an account](#)

### Hide word cloud

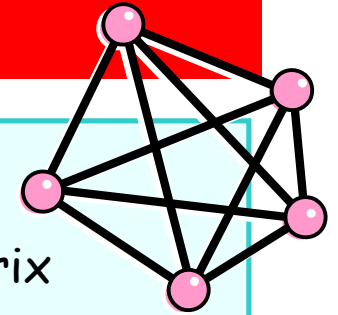
sorrow hesitation hysteria hubris fret hardship antagonism withdrawal **fight threat** upheaval peril rage indignation unhappiness phobia unease hatred urgency nagging aggression angst danger abuse aggravation burden revulsion resistance crushing inferiority suicide intolerance frustration confusion disturbance chaos discrimination vexation **death** negativity apprehension insecurity blood tedium depression hunger struggle silence overwhelming instability crippling **conflict** vulnerability dislike sweat turmoil shame resignation harm strain deprivation rejection pain ignorance oppression paranoia trouble outrage **crisis** trial pity stigma embarrassment violence paralysis emptiness **debt** anxiety privation prejudice seizure disillusionment injury **attack** horror indifference cancer **fear** stressful hate ego envy remorse terror disruption difficulty fever fatigue grief dismay suspicion debilitating pessimism gloom scepticism darkness guilt misery torture unrest impotence stress doubt **opposition** friction displeasure criticism humiliation weakness failure intimidation disorder panic animosity anger hostility inconvenience cynicism **loss** caution disbelief protest desperation disease aversion **war** agony resentment bitterness **risk** disappointment consternation inability poverty isolation illness migraine repression disability addiction distress uncertainty mistrust denial despair reluctance trauma skepticism harassment regret **problem** illusion dependence misunderstanding shock tears distrust noise alarm headache worry **tension**

⇒ [Rival Koreas gear up for talks this week to ease tension](#) - Reuters - 2011-02-06

Mood words: [conflict crisis war risk tension]

the crisis and restart international talks over the North's nuclear program. The two **Koreas** are still technically at war because an armistice not a treaty ended the 1950-53 Korean War. The meeting ...

## Conclusions: *Cliques all the way down ...*



- Cliques can be calculated in any adjacency graph (*but: NP-hard!*)

Analysis of large text corpora yields different kinds of adjacency matrix

- Cliques at different ontological levels can be mapped together

We can find cliques amongst *instances*, amongst *categories*, and amongst *domains*

- Analogy within an Ontology can be viewed as a clique-alignment problem

Mapping instance cliques → *category cliques* gives implicit (unlabelled) categories

- Analogical Cliques enrich Ontological Structure

[Afflatus.ucd.ie/dorian](http://Afflatus.ucd.ie/dorian)

E.g., supports intelligent (guided) exploration of large idea spaces (creative IR)