RESTAURANTS ANALYTICS AND THE DUTCH MOVIE WORLD

JUST SOME HOBBY PROJECTS

DATAVIZ MEETUP
10-10-2019
AMSTERDAM

Longhow Lam
Freelance data scientist: Just contact me if you need me :-)

Let's link on Linkedin ☺

https://www.linkedin.com/in/longhowlam

LHL
DSD

AGENDA

- INTRODUCTION
- RESTAURANTS ANALYTICS,
- DUTCH FILM WORLD

(SOME TEXT MINING, ASSOCIATION RULES MINING AND GRAPH ANALYSIS)

Maybe there is time to share two statistics in a small experiment

INTRODUCTION

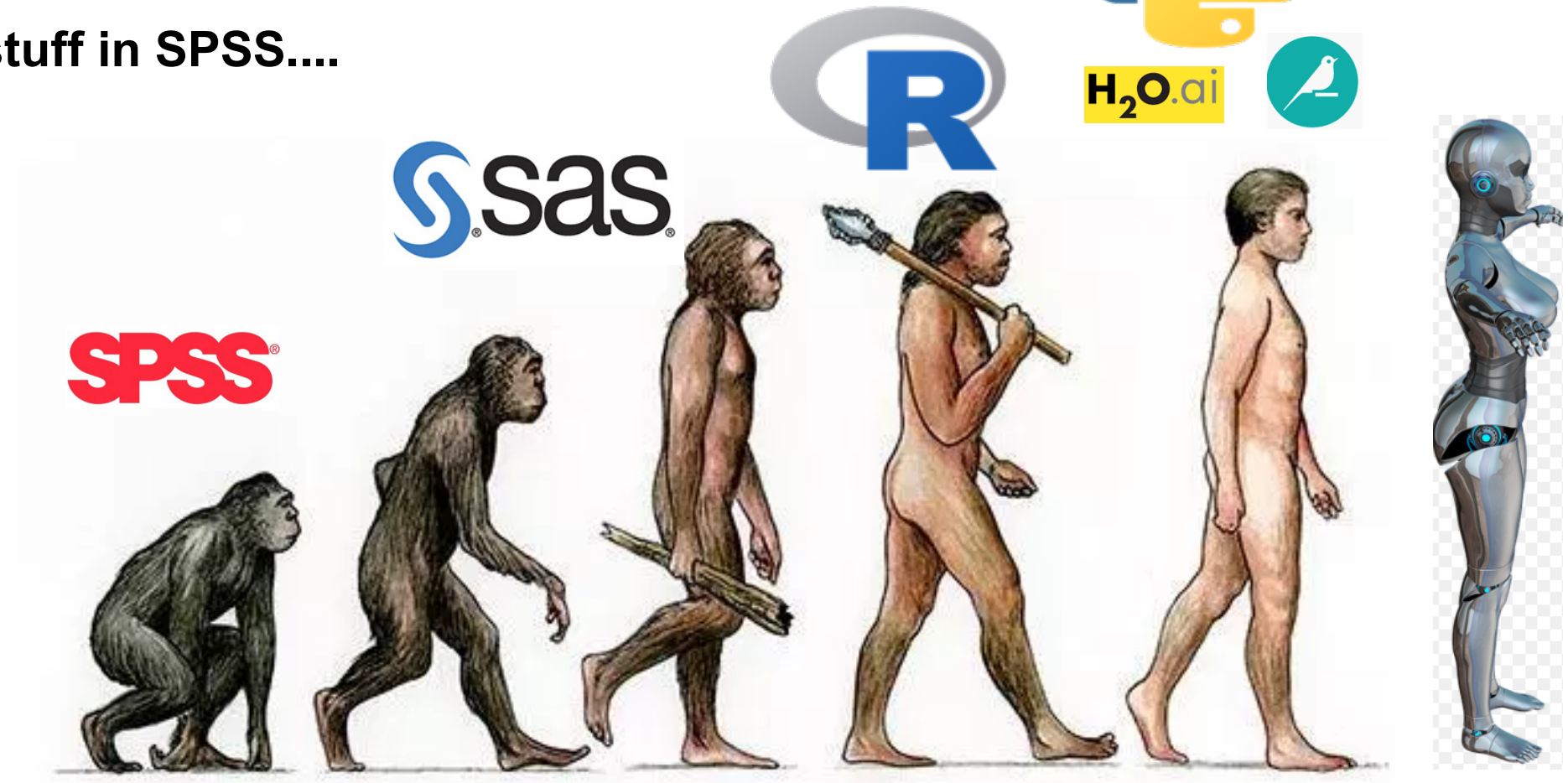
- An overview of different data science (machine learning) techniques
- Applied on some ‘playfull’ hobby projects
- I can never disclose company data in public → Scrape data
- However, all techniques are applied in “*real life*”
- My data science tool set:

INTRODUCTION

But you need to learn your whole life!
Data science environments they evolve, come and go

??

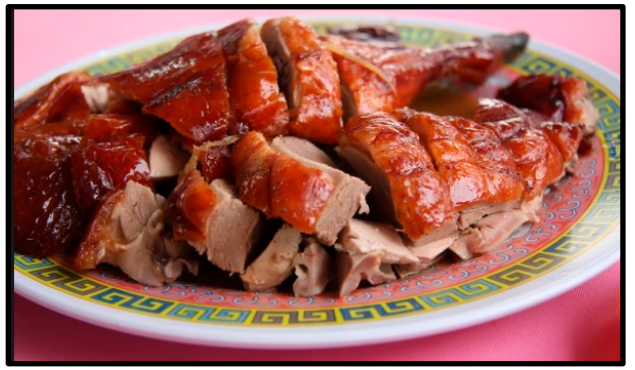
I once did stuff in SPSS....



Advanced Restaurant Analytics

LHL
DSD

RESTAURANT ANALYTICS



Business pain

- What are the key words for a good / bad restaturants?
- I have eaten Chinese, OK nice! But where to eat the next time?

Approach

Look at restaurant reviews and look where reviewers went

reviewer	RestaurantNaam	Review	keuken	datum	eten	service	decor
marian_groot	Het Ei Van Columbus	Heerlijk gegeten! Vooraf een heerlijke proeverij w...	FRANS	23/06/2013	8	9	8
Kremers	Het Ei Van Columbus	Vanwege verjaardag van Oma uit eten. Keuzemenu bes...	FRANS	15/06/2013	8	8	7
Mark-Gerards	Het Ei Van Columbus	Wij hebben hier met 6 personen gegeten en het is o...	FRANS	29/12/2012	10	10	1
posth151	Het Ei Van Columbus	De bediening was zonder meer vriendelijk en gastvr...	FRANS	28/12/2012	7	8	6
CulinairOegstgeest	Het Ei Van Columbus	Lekker gegeten in dit restaurant op mooie lokatie ...	FRANS	09/11/2012	7	7	7

Good and bad words

- Simpel ***non deep learning*** text mining apporach
- We now have a binairy Target: **BAD** = “score < 5” **GOOD** = “score > 8”
- Train a regularized logistic regression and look at largest and smallest coefficients

Review	aardig	eten	italiaans	Keuken	…	…	vis	zout	Target
Review 1	1	0	0	0	…	…	1	1	BAD
Review 2	0	1	0	1			0	1	BAD
…									…
…									…
…									…
Review N	0	0	1	1	…	…	1	0	GOOD

There is a problem, this matrix is very sparse and there are (too) many columns

TARGET PREDICTION WITH LASSO LOGISTIC REGRESSION

TERM DOCUMENT MATRIX

For each term (column) estimate a β (beta) parameter

Too many columns for regression: regularization is needed!

For example: “lasso” regression

$$\beta^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=i}^{N} \left\{ \log\left(1 + e^{\beta_0 + xi\beta}\right) - yi(\beta_0 + xi\beta) \right\}$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

Define target and Count vectorizer

In [3]:
```
###  take the bad and good reviews
review_sample_modeling = (
    review_sample
    .query('eten < 5 | eten > 8')
)

### create binary target, bad if review score is < 6 and review score > 8
review_sample_modeling = review_sample_modeling.assign(target = np.where(review_sample_modeling.eten < 6,1,0))
```

In [4]:
```
review_sample_modeling.target.value_counts()
```

Out[4]:
```
0    8507
1    1067
Name: target, dtype: int64
```

In [5]:
```
cv = CountVectorizer(ngram_range=(1,2))
cv.fit(review_sample_modeling.cleaned_review)
X = cv.transform(review_sample_modeling.cleaned_review)
target = review_sample_modeling.target

X_train, X_val, y_train, y_val = train_test_split(
    X, target, train_size = 0.75
)
```

In [6]:
```
X_train.shape
```

Out[6]:
```
(7180, 336228)
```

Logistic regression

In [94]:
```
for c in [0.01, 0.05, 0.25, 0.5, 0.75, 0.95, 1]:
    lr = LogisticRegression(C=c)
    lr.fit(X_train, y_train)
    print (
        "Accuracy for C=%s: %s"
        % (c, accuracy_score(y_val, lr.predict(X_val)))
    )
```

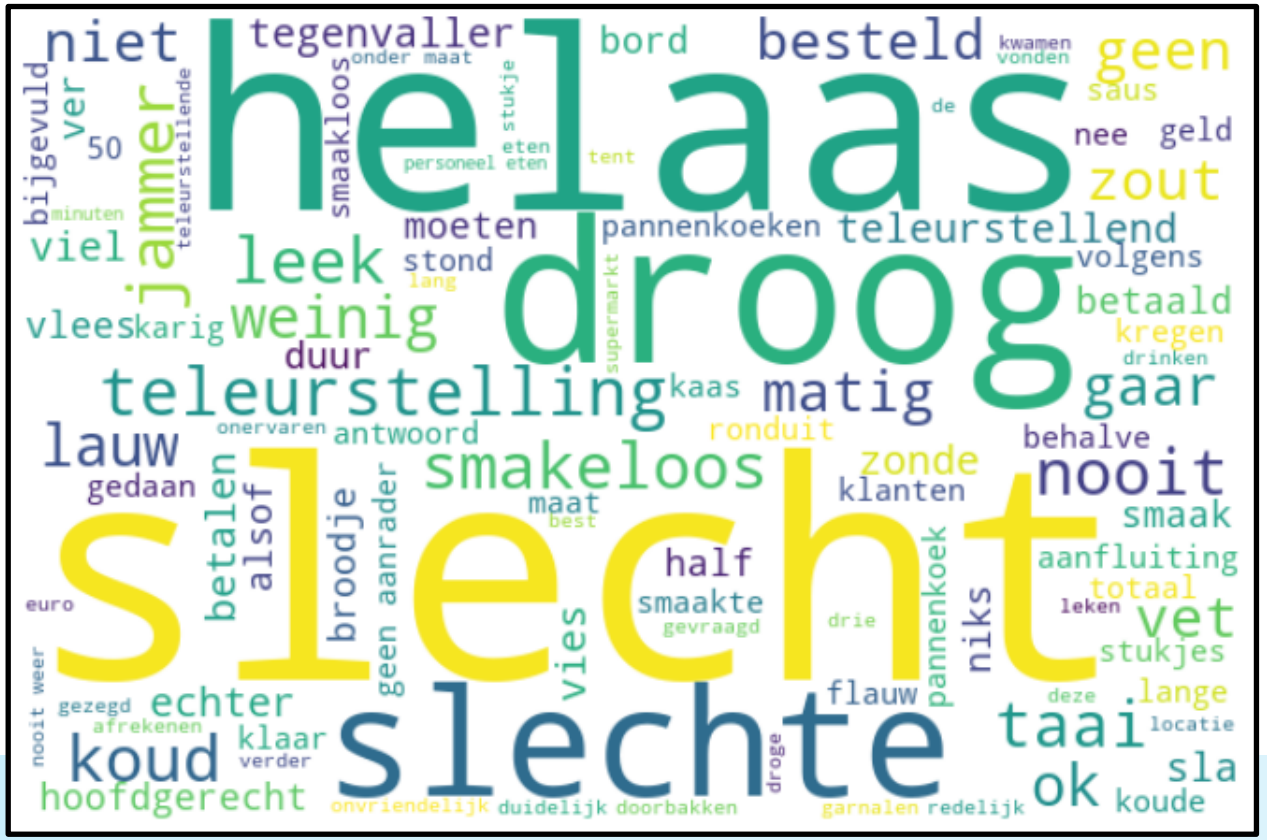
```
/home/longhowlam/anaconda3/lib/python3.7/site-packages/sklearn/linear_model/logistic.py:432: FutureWarning: Default solver will be changed to 'lbfg
  FutureWarning)

Accuracy for C=0.01: 0.9486215538847118
Accuracy for C=0.05: 0.9670008354218881
Accuracy for C=0.25: 0.9736842105263158
Accuracy for C=0.5: 0.9741019214703425
Accuracy for C=0.75: 0.9741019214703425
Accuracy for C=0.95: 0.9736842105263158
Accuracy for C=1: 0.9732664995822891
```

Good and bad words

size of word is related to β estimates

Reviews data and Jupyter notebook with analysis on [my GitHub]



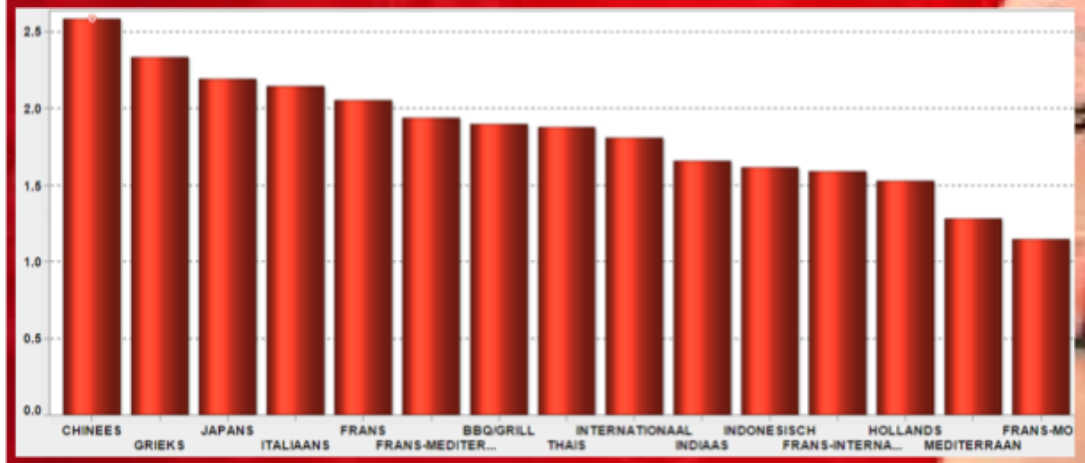
A FEW FACTS… | IENS DATA (TRADITIONAL BI)

Most occuring restaurant name (39 times)

% Sustainable kitchens

Biological	(67%)
French	(58%)
Fish	(44%)
Vegetarian	(39%)
…	
…	
…	
Chinese	**(3%)**

700 reviews on a "normal" Satuday
Valentine 2015 had 1200 reviews (1.7 times)



12 times

23 times

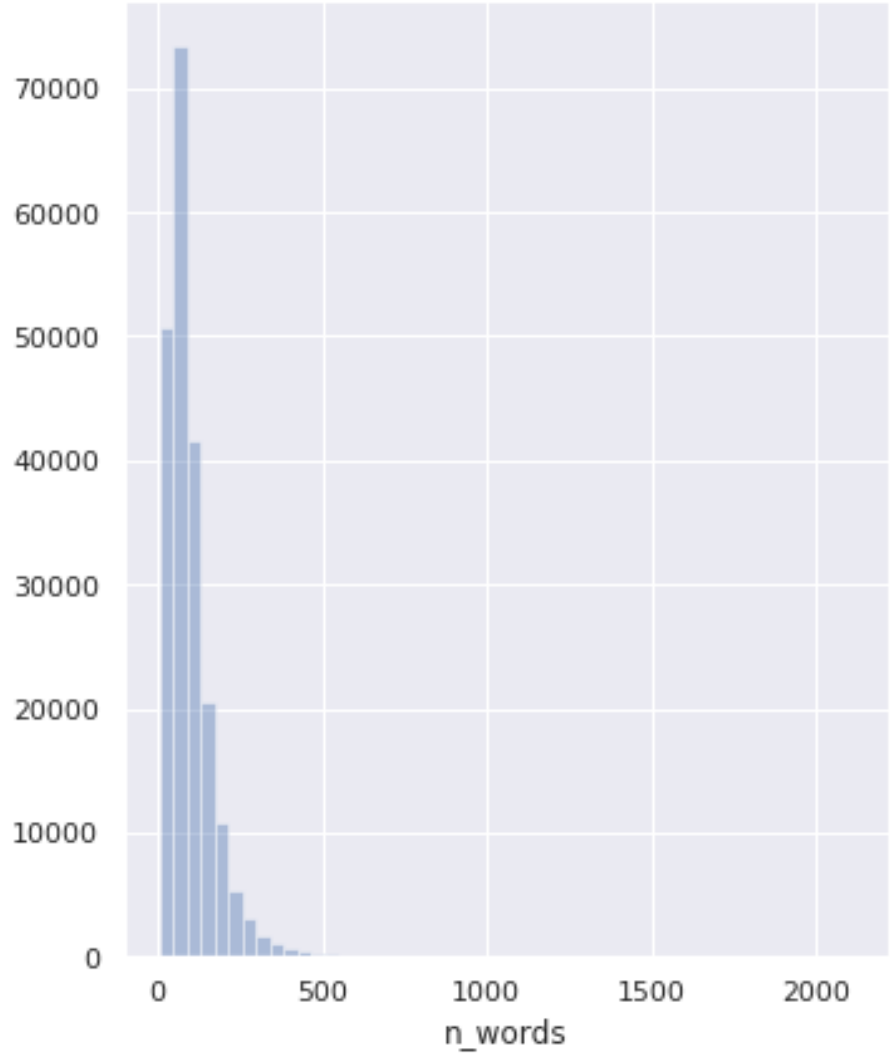
RHODOS
Grieks Restaurant

De Herberg
RESTAURANT

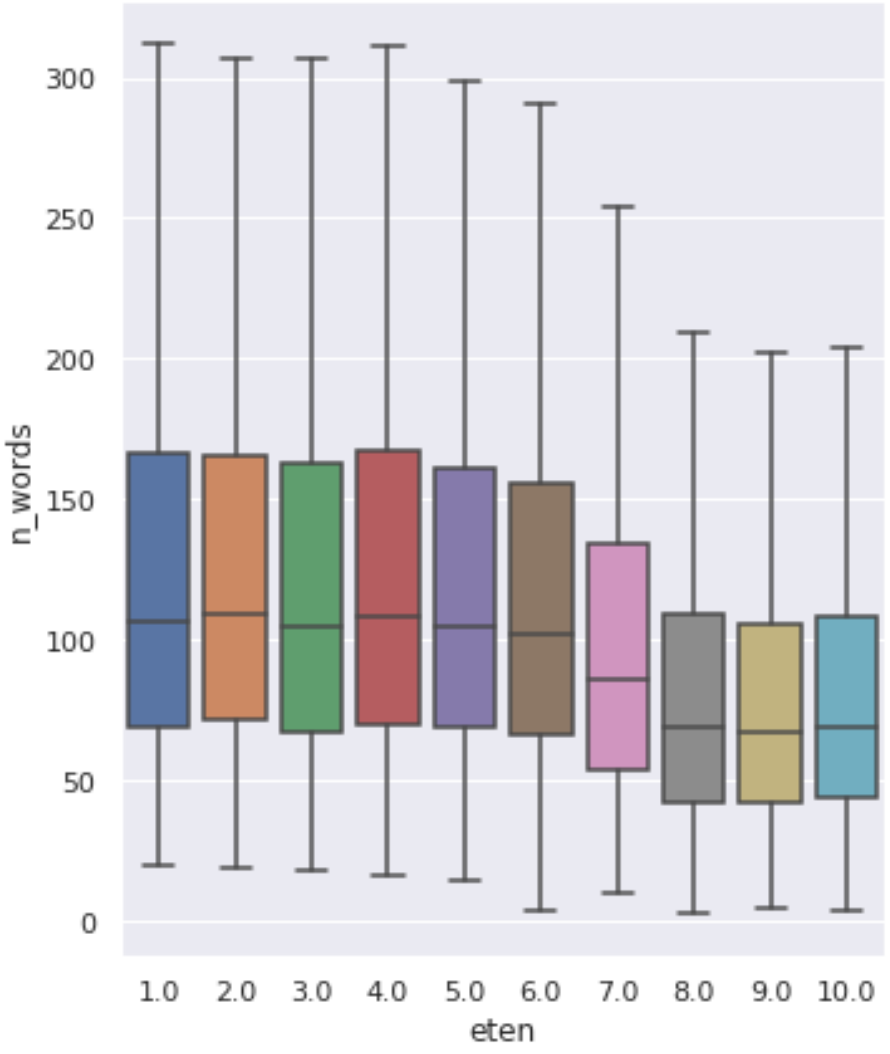
Among Dutch restaurants (6 keer)

A FEW FACTS... | IENS DATA

Distribution of number of words in a review



Lower scores observed at longer reviews



ASSOCIATION RULES MINING

ALSO CALLED MARKET BASKET ANALYSIS

Identify **frequent item sets** (rules) in transactional data:

```
IF items A and B THEN item C   {A, B} → {C}
IF items X THEN item Y and Z   {X} → {Y, Z}
```

When is rule frequent? If the '***support***' > a treshold

Support

$$\text{Support } \{X \rightarrow Y\} = \frac{\#\text{ trxs. } \{X \rightarrow Y\}}{\text{Total \# trxs.}}$$

Chips –> Milk 0.002%

Chips –> Beer 0.823%

ASSOCIATION RULES MINING

ALSO CALLED MARKET BASKET ANALYSIS

Other statistics used to assess the usefulness of a rule

Lift & Confidence

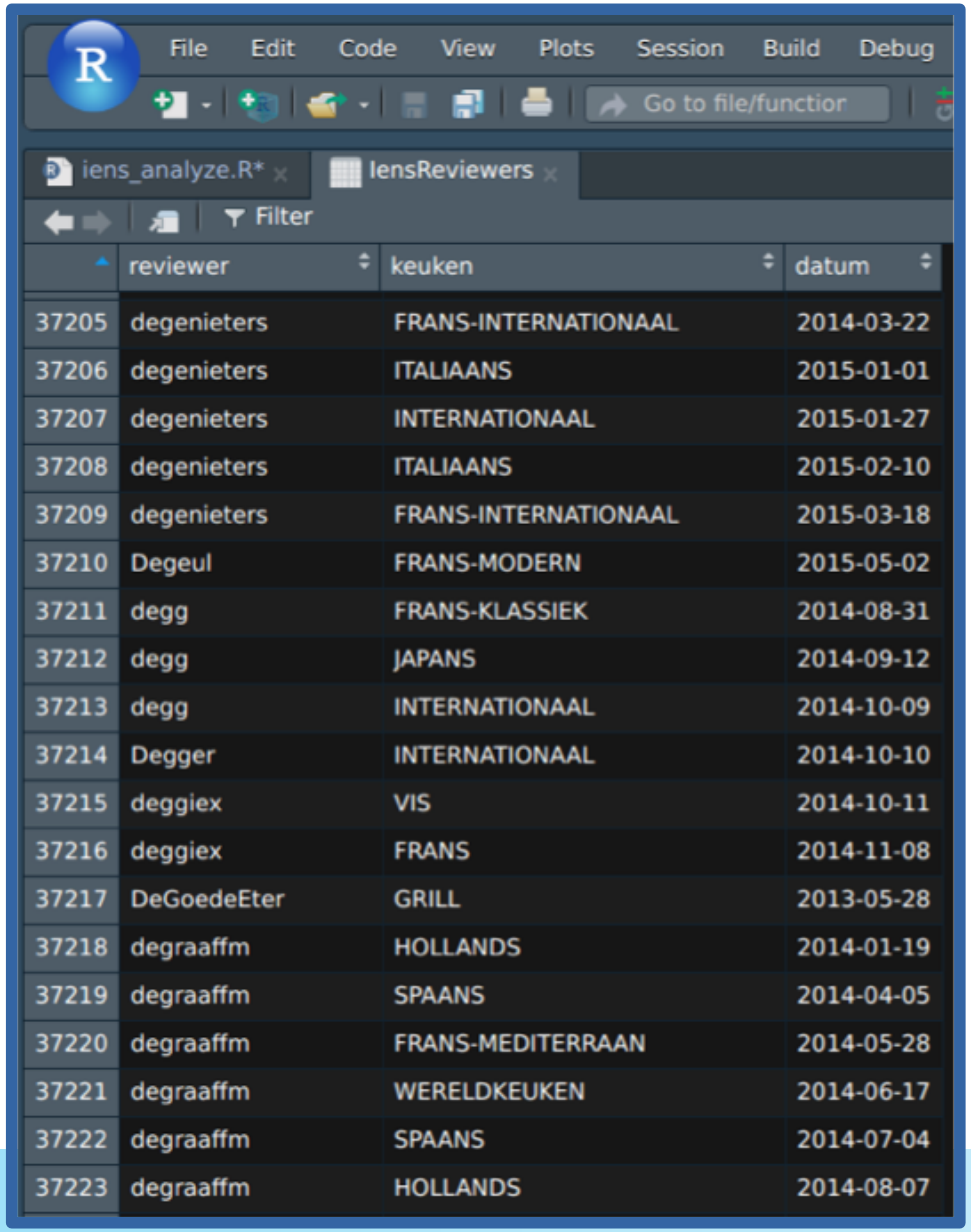
$$\text{Lift}\{X \longrightarrow Y\} = \frac{\text{Support}\{X \rightarrow Y\}}{\text{Support}(X) \ * \ \text{Support}(Y)}$$

$$\text{Conf}\{X \longrightarrow Y\} = \frac{\text{Support}\{X \rightarrow Y\}}{\text{Support}(X)}$$

Example:

a lift van 8.3 for {Chips} → {Beer} means

If I know someone has already bought Chips then it is 8.3 more likely that he will also buy beer



	reviewer	keuken	datum
37205	degenieters	FRANS-INTERNATIONAAL	2014-03-22
37206	degenieters	ITALIAANS	2015-01-01
37207	degenieters	INTERNATIONAAL	2015-01-27
37208	degenieters	ITALIAANS	2015-02-10
37209	degenieters	FRANS-INTERNATIONAAL	2015-03-18
37210	Degeul	FRANS-MODERN	2015-05-02
37211	degg	FRANS-KLASSIEK	2014-08-31
37212	degg	JAPANS	2014-09-12
37213	degg	INTERNATIONAAL	2014-10-09
37214	Degger	INTERNATIONAAL	2014-10-10
37215	deggiex	VIS	2014-10-11
37216	deggiex	FRANS	2014-11-08
37217	DeGoedeEter	GRILL	2013-05-28
37218	degraaffm	HOLLANDS	2014-01-19
37219	degraaffm	SPAANS	2014-04-05
37220	degraaffm	FRANS-MEDITERRAAN	2014-05-28
37221	degraaffm	WERELDKEUKEN	2014-06-17
37222	degraaffm	SPAANS	2014-07-04
37223	degraaffm	HOLLANDS	2014-08-07

A little bit of R code

```
16  ## generate market basket rules
17  rules <- apriori(
18    ienstrx,
19    parameter = list(
20      supp = 0.00020,
21      conf = 0.18,
22      maxlen = 2
23    )
24  )
25
26  ## eerste tien regels op basis van lft
27  inspect( sort(rules, by = "lift")[1:10])
28
```

IENS RESTAURANT

LENGTH TWO RULES A → B

```
> inspect( sort(rules, by = "lift")[2:20])
     lhs                rhs              support     confidence lift     count
[1]  {MALEISISCH}    => {THAIS}          0.0001447513 0.2542373  7.845620  15
[2]  {KOREAANS}      => {CHINEES}        0.0007527068 0.2795699  7.079841  78
[3]  {KANTONEES}     => {CHINEES}        0.0007913072 0.2789116  7.063170  82
[4]  {AFRIKAANS}     => {THAIS}          0.0001544014 0.2253521  6.954240  16
[5]  {KOREAANS}      => {JAPANS}         0.0008492077 0.3154122  6.760063  88
[6]  {VIETNAMEES}    => {CHINEES}        0.0013896126 0.2622951  6.642373 144
[7]  {VIETNAMEES}    => {THAIS}          0.0011194102 0.2112933  6.520392 116
[8]  {BALKAN}        => {GRIEKS}         0.0002316021 0.2400000  6.291485  24
[9]  {MALEISISCH}    => {JAPANS}         0.0001640515 0.2881356  6.175458  17
[10] {AFRIKAANS}     => {INDIAAS}        0.0001254511 0.1830986  6.096971  13
[11] {DUURZAAM}      => {BIOLOGISCH}     0.0003377531 0.1891892  6.004569  35
[12] {ZWITSERS}      => {MEDITERRAAN}    0.0001158011 0.3428571  6.003534  12
[13] {KOREAANS}      => {THAIS}          0.0005114547 0.1899642  5.862188  53
[14] {WOKKEN}        => {CHINEES}        0.0008492077 0.2303665  5.833812  88
[15] {VIETNAMEES}    => {JAPANS}         0.0014089128 0.2659381  5.699710 146
[16] {SCHOTS}        => {MEDITERRAAN}    0.0001254511 0.3250000  5.690850  13
[17] {SURINAAMS}     => {CHINEES}        0.0012448613 0.2216495  5.613062 129
[18] {TIBETAANS}     => {CHINEES}        0.0002219520 0.2211538  5.600510  23
[19] {DUURZAAM}      => {BBQ/GRILL}      0.0003281030 0.1837838  5.245051  34
```

Very generic rules
Lift is not really high

[Interactieve netwerk]

IENS RESTAURANT | LENGTH THREE RULES A, B ⟶ C

Much more specific, higher lift

```
> inspect( sort(rules, by = "lift")[2:200])
      lhs                                  rhs                  support      confidence lift     count
[1]   {KOREAANS,VIS}                    => {VIETNAMEES}         0.0001061510 0.3548387  66.97726 11
[2]   {KOREAANS,THAIS}                  => {VIETNAMEES}         0.0001544014 0.3018868  56.98237 16
[3]   {CHINEES,KOREAANS}                => {VIETNAMEES}         0.0002123019 0.2820513  53.23834 22
[4]   {CHINEES,ETHIOPISCH}              => {VIETNAMEES}         0.0001061510 0.2750000  51.90738 11
[5]   {BBQ/GRILL,KOREAANS}              => {VIETNAMEES}         0.0001158011 0.2727273  51.47839 12
[6]   {INDONESISCH,KOREAANS}            => {VIETNAMEES}         0.0001158011 0.2666667  50.33443 12
[7]   {JAPANS,KOREAANS}                 => {VIETNAMEES}         0.0002123019 0.2500000  47.18852 22
[8]   {INTERNATIONAAL,KOREAANS}         => {VIETNAMEES}         0.0002509023 0.2203390  41.58989 26
[9]   {CHINEES,VEGETARISCH}             => {VIETNAMEES}         0.0001833517 0.2111111  39.84809 19
[10]  {BBQ/GRILL,SURINAAMS}             => {VIETNAMEES}         0.0001158011 0.2033898  38.39066 12
[11]  {ENGELS,MEDITERRAAN}              => {VIETNAMEES}         0.0001254511 0.2031250  38.34068 13
[12]  {ITALIAANS,KOREAANS}              => {VIETNAMEES}         0.0001833517 0.1958763  36.97245 19
[13]  {LIBANEES,MEDITERRAAN}            => {VIETNAMEES}         0.0001158011 0.1875000  35.39139 12
[14]  {INDIAAS,VEGETARISCH}             => {SURINAAMS}          0.0001351012 0.1917808  34.14687 14
[15]  {ETHIOPISCH,HOLLANDS}             => {VEGETARISCH}        0.0001351012 0.2641509  32.50939 14
[16]  {BIOLOGISCH,SURINAAMS}            => {VEGETARISCH}        0.0001544014 0.2580645  31.76032 16
[17]  {ETHIOPISCH,HOLLANDS}             => {PIZZERIA}           0.0001254511 0.2452830  30.58688 13
[18]  {ETHIOPISCH,MEDITERRAAN}          => {VEGETARISCH}        0.0001158011 0.2352941  28.95794 12
[19]  {TAPAS/MEZZE,VIETNAMEES}          => {BELGISCH}           0.0001158011 0.2790698  28.32408 12
[20]  {ETHIOPISCH,ITALIAANS}            => {VEGETARISCH}        0.0001544014 0.2222222  27.34917 16
[21]  {ETHIOPISCH,FRANS-INTERNATIONAAL} => {VEGETARISCH}        0.0001158011 0.2181818  26.85191 12
[22]  {ETHIOPISCH,FRANS}                => {VEGETARISCH}        0.0001544014 0.2133333  26.25520 16
```

Interactief plaatje

IENS RESTAURANT

VIRTUAL ITEMS: MAKE IT EVEN MORE PERSONAL

Transaction data with customers and items

klant	ITEM
1	A
1	X
2	A
2	B
2	C
3	E
3	T
4	S

possible rules

{ A, B } → { C }

{ X } → { Z }

Add customer features as virtual items

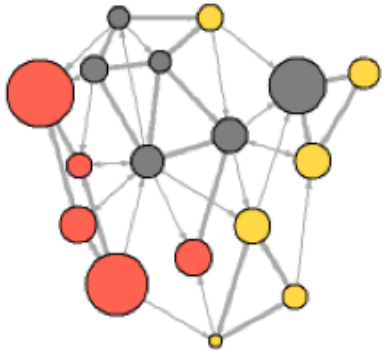
klant	ITEM
1	A
1	X
1	**Male**
1	**(18, 25]**
2	A
2	B
2	C
2	**Male**
2	**(45, 65]**
3	E
3	T
3	**Male**
4	**(30, 35]**
4	S
4	**Male**
4	**(30, 35]**

possible rules

{ Male, (18, 25], A, B } → { C }

{ Female, (40,45], X } → { Z }

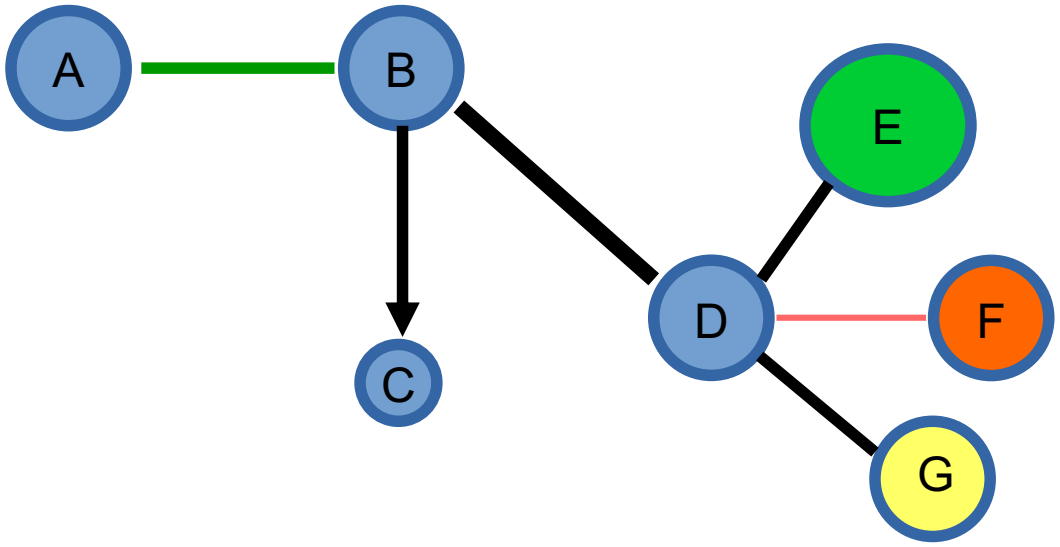
The Dutch movie world in a graph



LHL
DSD

Node or Vertex a point in the network

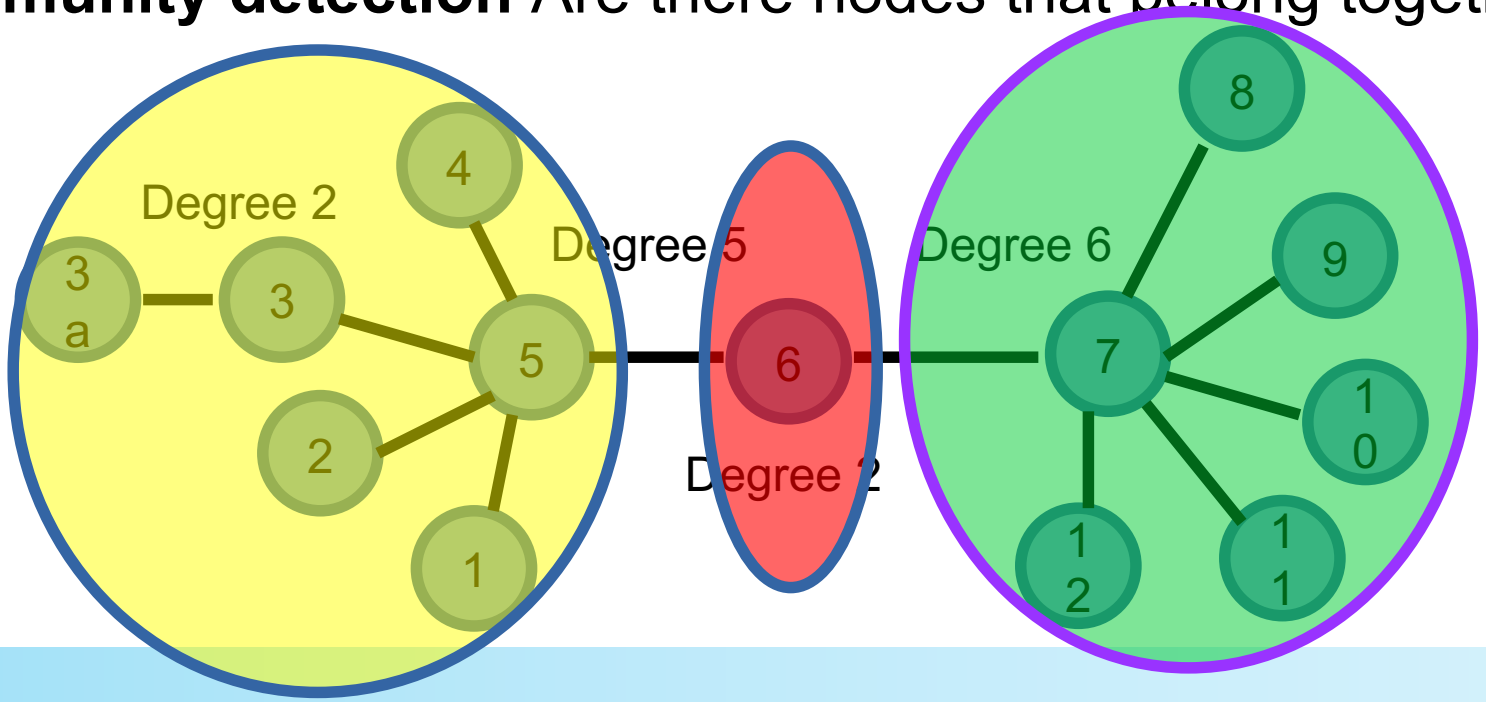
Edge or Link a relation between two nodes (can be directional)



Node Centraliteit How central is a node

* Degree (number of connections)
* Betweenness (number of shortest paths through a node)
* Eigencentrality (Google's page rank is a version of this)

Community detection Are there nodes that belong together?



Node 6 and 3 have the same **Degree**,

But node 6 has a higher **Betweennes** than node 3

WWW.IMDB.COM | **INTERNET MOVIE DATABSE**

Download movie data:

* Dutch movies in the last 25 years
* Per movie we know the cast and crew
* A node is a persoon
* Node ***X*** links with node ***Y*** if ***X*** and ***Y*** were in the same movie

R Script of this analysis see [my GitHub Repo]

DUTCH MOVIE | WORLD IN A NETWORK GRAPH

nodes
Chantal Jantzen
`Hans de Wolf
…

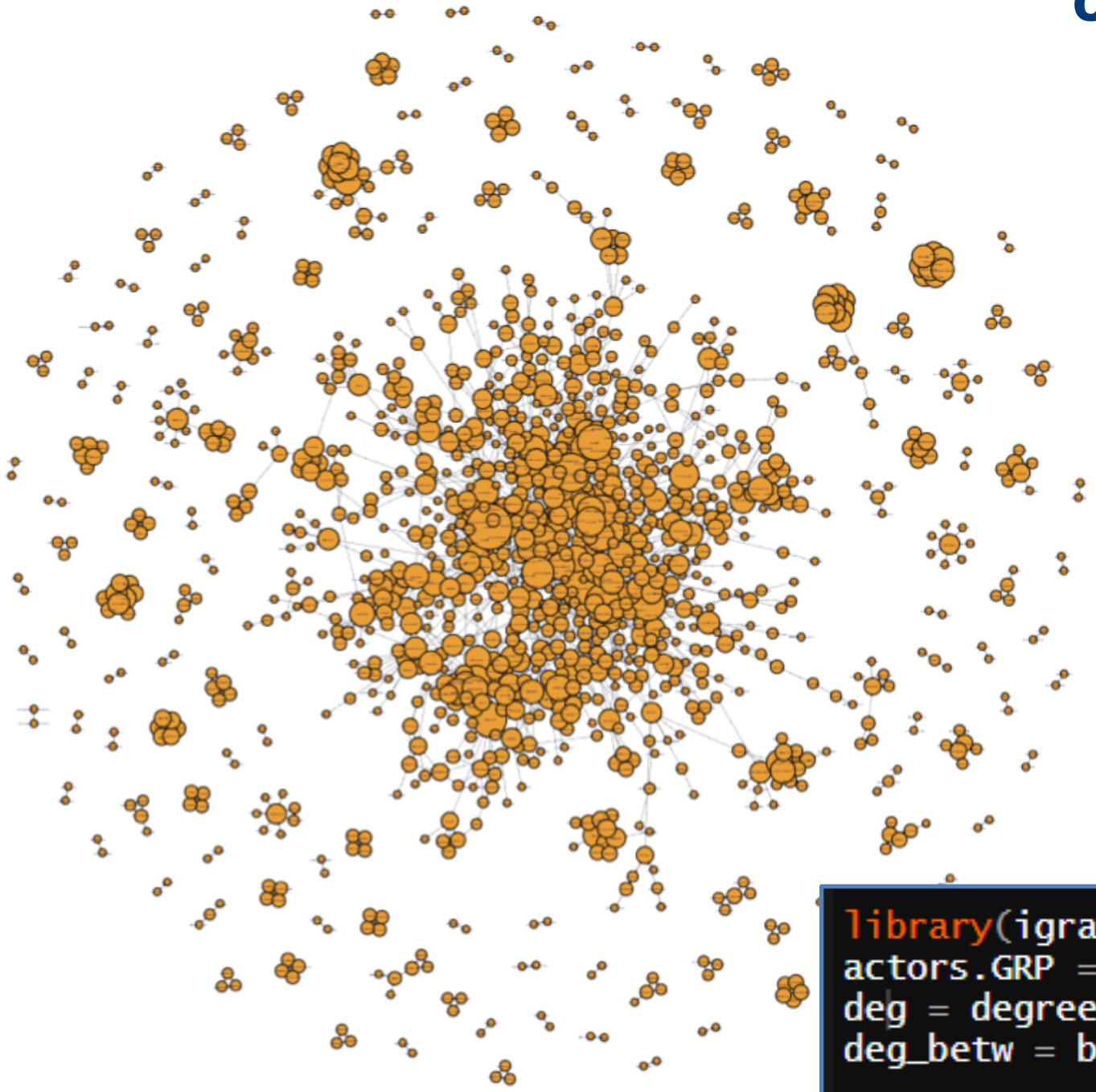
edges	
Chantal Jantzen	Stef Tijding
Hans de wolf	Jeroen Krabee
Johan Nijenhuis	Frans van Gestel
…	…..

Actors are in the nodes data.frame
Their links are in the edges data.frame

```
visNetwork(nodes, edges)
```

Interactive graph

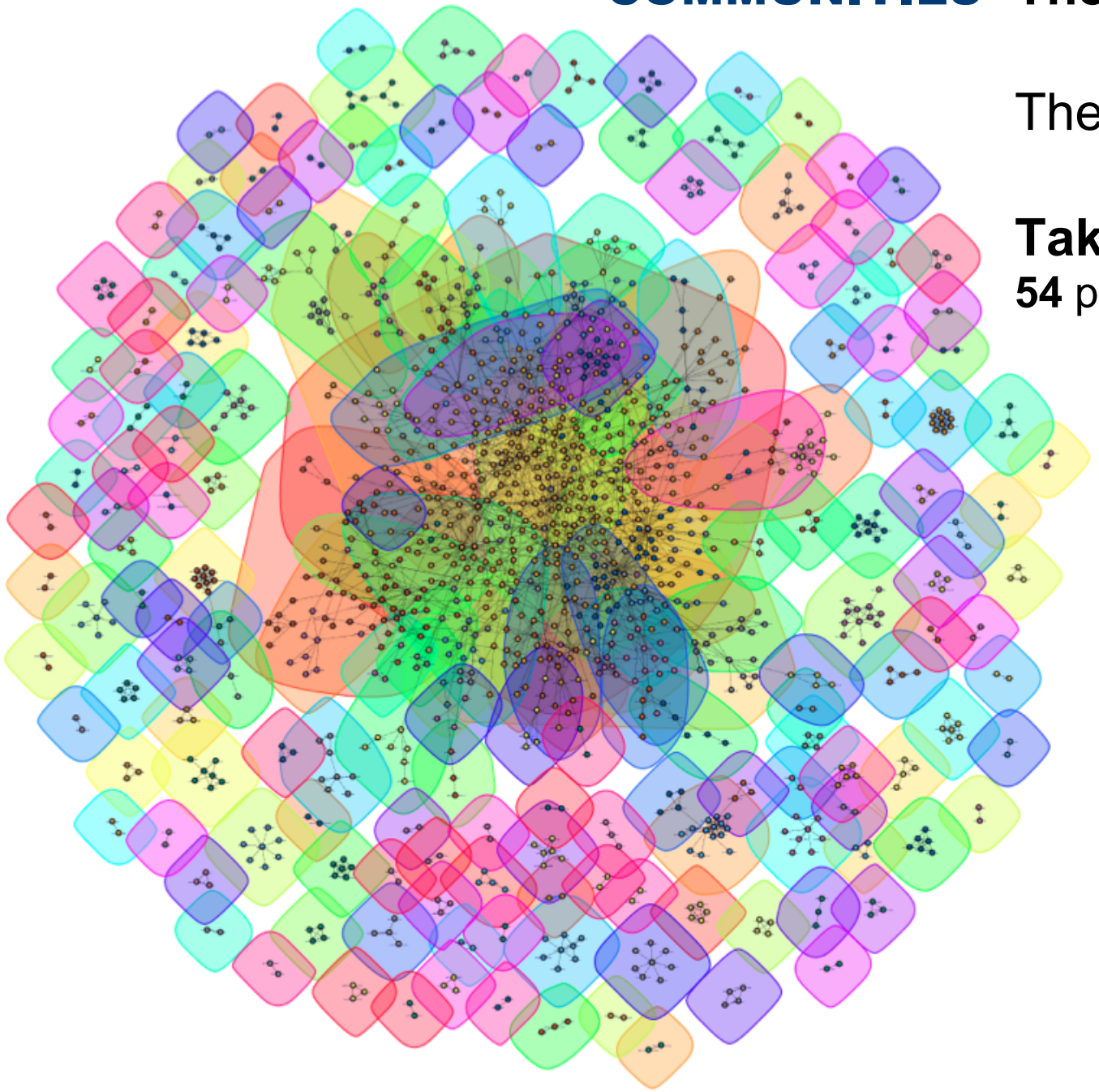
CENTRALITY



persoon	centrality
Paul Ruven	74
Fons Merkies	68
Frans van Gestel	60
Pieter van Huystee	52
Hanneke Niens	50
Stef Tijdink	44
Alain De Levita	42
Frank van den Engel	42
Jean-Pierre Claes	42
Jeroen Beker	42
Hans de Wolf	36
Paul M. van Brugge	36
Ton Peters	36
Janneke Doolaard	32
Gregor Meerman	30
Jacko van 't Hof	30
Jan van der Zanden	30
Johan Nijenhuis	30
Martijn van Nellestijn	30
Peter Brugman	30

```
library(igraph)
actors.GRP = graph_from_data_frame(edges)
deg = degree(actors.GRP, mode="all")
deg_betw = betweenness(actors.GRP,directed = FALSE)
```

COMMUNITIES



There are 1257 persons

They are divided in 191 community's

Take community 6:

54 persons in a wordcloud (Centrality based)

A LITTLE STATISTICAL EXPERIMENT

Can you shake hands with the two neighbors?

Two statistics I like to share with you:

A LITTLE STATISTICAL EXPERIMENT

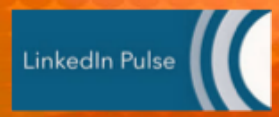
50.1% of all people don't wash there hands after using the toilet

A LITTLE STATISTICAL EXPERIMENT

84.6% **of all statistics are made up at the spot!!**

Thanks for your time! Questions?

Need me as Freelancer? Let's have a cup of coffee



@longhowlam

https://longhowlam.wordpress.com/

https://www.linkedin.com/in/longhowlam

LHL
DSD