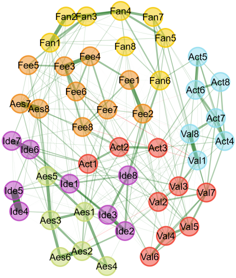
Dynamic Exploratory Graph Analysis

DS-5740 Advanced Statistics



Overview: Week 12

Goals for the Week

- Understand (intensive) longitudinal measurement
- Learn how to use and apply dynamic exploratory graph analysis
- Uncover how to detect clusters of people in dynamic data

Dynamic Exploratory Graph Analysis

Time series is back!

Recall: Types

- **cross-sectional**: measurement at a single time point (a *cross-section* in time)
- **panel**: measurement at multiple single time points (usually equally spaced in time)
- **longitudinal**: multiple measurements across time (usually much more than panel) that can be on the order of minutes, hours, days, weeks, months, or years

Dynamic Exploratory Graph Analysis

longitudinal: multiple measurements across time (usually much more than panel) that can be on the order of minutes, hours, days, weeks, months, or years

- Minutes, hours, days: “intensive”
- Weeks, months, years: “standard”

Intensive longitudinal data is most often used to capture dynamics across a short time window for processes that tend to have more rapid shifts from moment-to-moment

For this reason, often referred to as *ecological momentary assessment* (EMA)

Recall our example of emotions during the pandemic…

Table 1. Ecological Momentary Assessment Items, Queried Four Times per Day Over 2 Weeks

No.	Abbreviation	Item	Change	*p*
1	Relax	I found it difficult to relax	–0.11	.00
2	Irritable	I felt (very) irritable	–0.08	.00
3	Worry	I was worried about different things	–0.12	.00
4	Nervous	I felt nervous, anxious, or on edge	–0.13	.00
5	Future	I felt that I had nothing to look forward	–0.05	.00
6	Anhedonia	I couldn't seem to experience any positive feeling at all	–0.03	.07
7	Tired	I felt tired	–0.05	.00
8	Alone	I felt like I lack companionship, or that I am not close to people	–0.04	.02
9	Social_offline	I spent __ on meaningful, offline, social interaction	–0.02	.14
10	Social_online	I spent __ using social media to kill/pass the time	–0.06	.00
11	Outdoors	I spent __ outside (outdoors)	–0.03	.08
12	C19_occupied	I spent __ occupied with the coronavirus (e.g., watching news, thinking about it, talking to friends about it)	–0.18	.00
13	C19_worry	I spent __ thinking about my own health or that of my close friends and family members regarding the coronavirus	–0.16	.00
14	Home	I spent __ at home (including the home of parents/partner)	0.03	.03

Note: All items had five answer options. Items 1 through 8: 1 = *not at all*, 2 = *slightly*, 3 = *moderately*, 4 = *very*, 5 = *extremely*. Items 9 through 14: 1 = *0 min*, 2 = *1–15 min*, 3 = *15–60 min*, 4 = *1–2 hr*, 5 = *>2 hr*. The "Change" column displays standardized coefficients of change from univariate regression models over the 54 assessment points, followed by *p* values for these changes.

Dynamic Exploratory Graph Analysis

What was the design?

- intensive longitudinal: 4 times per day for 2 weeks

What are the benefits?

- Real-time thoughts and feelings (no recollection)
- Captures dynamics (variability within and between people)

within-person: repeated measurements of an individual person

between-person: measurements collapsed *across* people

Variability

- Capture dynamics of variables
- Interested in...
  1. how variables change together
  2. whether variables "synchronize"
  3. whether individuals differ from one another and/or the sample

What models do we know that can capture variability in time series?

What models do we know that can capture variability in time series?

- TSLM: regression on an outcome
- Autoregression (AR): lagged outcome regressed on itself
- Vector autoregression (VAR): lagged variables regressed on each other
- (Generalized) ARCH: volatility of time series

Do any of these capture “how variables change together”?

What models do we know that can capture variability in time series?

- TSLM: regression on an outcome
- Autoregression (AR): lagged outcome regressed on itself
- Vector autoregression (VAR): lagged variables regressed on each other
- (Generalized) ARCH: volatility of time series

Do any of these capture "how variables change together"?

Vector autoregression

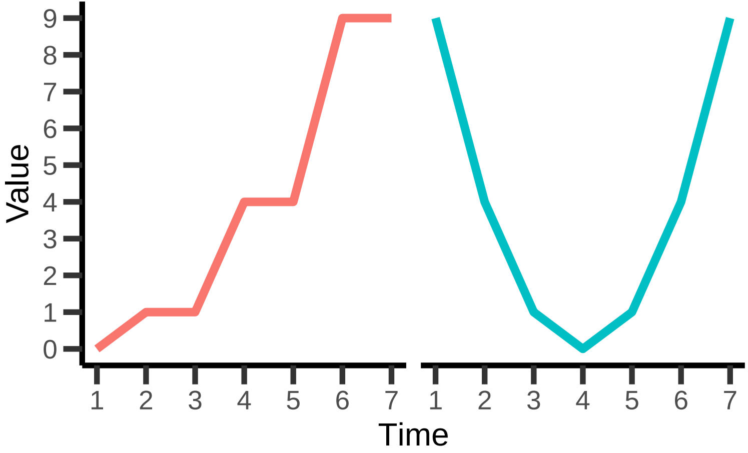
- common technique to look at how variables are changing together across time in many different fields

Time Considerations

- What is variability in time series data?

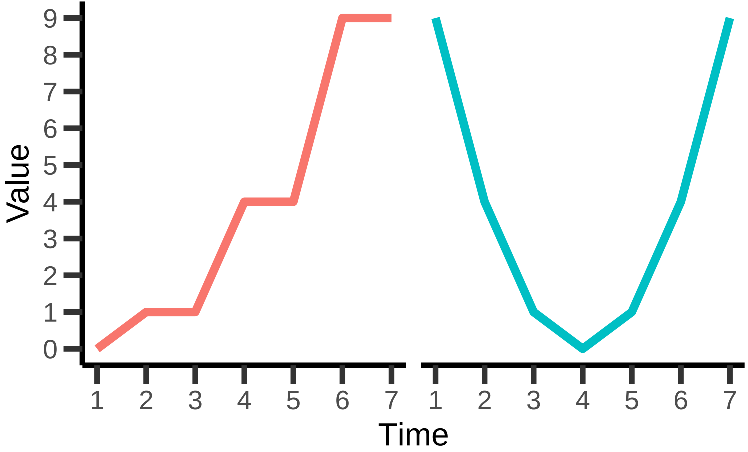
Time Considerations

- What is variability in time series data?



Time Considerations

These time series have the same variance (*SD* = 3.742)!



Time Considerations

- Variance of a time series does not capture its underlying dynamics

- This issue limits our ability to interpret *associations* between variables in our data

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

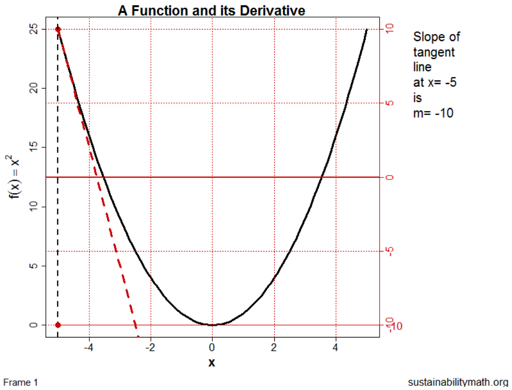
$$s^2 = \frac{\sum(y - \bar{y})^2}{n - 1}$$

Time Considerations

- Correlations with time:
  - red = 0.949
  - blue = 0
- Correlations with each other: 0.167

Time Considerations

- How can we capture the variability of the time series?



Time Considerations

- Differential equations: slopes (tangent lines) of curve
- First-order derivative: *velocity* (rate of change)
- Second-order derivative: *acceleration* (rate of rate of change)

Generalized Local Linear Approximation

- Integrals are computationally intensive
- Approximations are simpler, faster, and nearly as accurate

Generalized Local Linear Approximation

1. Create a time delay embedding
2. Compute average differences between values
3. Repeat for each sequence in embedding

Time Delay Embedding

```r
# Create time delay embedding
embedding <- Embed(
  x = df$y[df$value == "squared"], # univariate time series
  E = 3, # number of embedding columns
  tau = 1 # lag
)
```

E_1	E_2	E_3
9	4	1
4	1	0
1	0	1
0	1	4
1	4	9

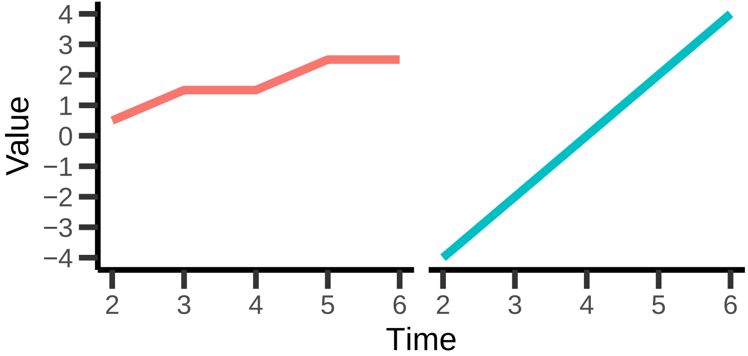
Derivatives

```r
# Compute derivatives
derivatives <- glla(
  x = df$y[df$value == "squared"], # univariate time series
  n.embed = 3, # number of embeddings
  tau = 1, # lag
  delta = 1, # time between observations
  order = 1 # order of derivative
)
```

Time	x	y	Moving Average	First Derivative
1	-3	9	NA	NA
2	-2	4	4.67	-4
3	-1	1	1.67	-2
4	0	0	0.67	0
5	1	1	1.67	2
6	2	4	4.67	4
7	3	9	NA	NA

Our Example

- These time series *do not* the same variance!



- Standard deviations
  - red = 0.837
  - blue = 3.162

Our Example

- Original relationship with time:
  - red = 0.949
  - blue = 0
- Derivative relationship with time:
  - red = 0.945
  - blue = 1
- Correlations with each other
  - Original = 0.167
  - Derivative = 0.945

What happened?

Original Time Series

- standard deviation: does not capture dynamics – it captures deviations from mean (time does not matter)

- correlation: only captures *linear* relationships

Original Time Series

- standard deviation: does not capture dynamics – it captures deviations from mean (time does not matter)

- correlation: only captures *linear* relationships

Derivative Time Series

- standard deviation: captures variability in how a variable *changes over time* (i.e., its dynamics)

- correlation: captures linear *and* nonlinear relationships

Interpretations

- Variance
  - low: small range of velocities (first-order derivatives) – there is little change over time
  - high: large range of velocities – there is lots of variability over time

Interpretations

- Variance
  - low: small range of velocities (first-order derivatives) – there is little change over time
  - high: large range of velocities – there is lots of variability over time
- Mean
  - positive ($>$ 0): generally increasing trend over time (i.e., changes tend to be more upward than downward)
  - negative ($<$ 0): generally decreasing trend over time (i.e., changes tend to be more downward than upward)
  - zero (0): increases and decreases *cancel* one another out

Dynamic Exploratory Graph Analysis

Dynamic Exploratory Graph Analysis

1. Compute Generalized Local Linear Approximation (GLLA) for each variable for *each* person's time series

2. Estimate EBICglasso across all people (stack each person's derivatives) and each individual person

3. Apply a community detection algorithm to the "population" network (all people) and "individual" networks (each person)

Empirical Example

- n = 122 completed the BFI-2
- Beeped 4 times a day for two weeks
- Completed around 10-15 Big Five Inventory 2 items at each beep
- Missing responses to non-queried items were *imputed*

Our Questions

- Do *variables* cluster into dimensions? Do we find the Big Five?
- Do *variables* cluster into the same dimensions for each person?
- Do *people* cluster into sub-groups or *types*?

Do *variables* cluster into dimensions? Do we find the Big Five?

- Load {EGAnet} and data

```
# Load {EGAnet}
library(EGAnet)

# Load data
load("../data/esm_data.RData")
```

- Length of each time series

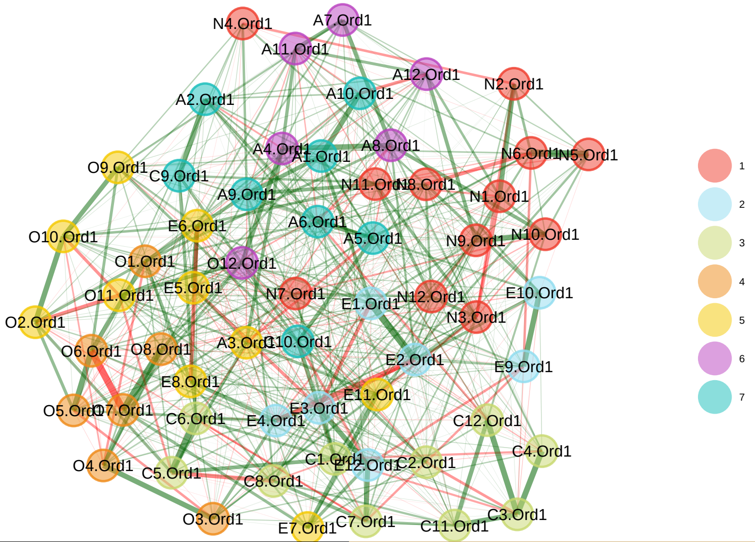
```
# Length of each time series
table(esm_data$ID)
```

Do *variables* cluster into dimensions? Do we find the Big Five?

```
# Estimate Dynamic EGA
bfi2_dynamic <- dynEGA(
  data = esm_data, # long format dataset
  n.embed = 4, # number of GLLA embeddings (4 beeps a day)
  delta = 1, # lag = 1
  level = c("population", "individual"),
  # population and individual networks
  id = 1, # first column
  use.derivatives = 1, # first order derivatives
  model = "glasso", # estimate Gaussian graphical model
  algorithm = "louvain" # community detection algorithm
)
```

Dynamic Exploratory Graph Analysis

```r
# Plot population network
plot(bfi2_dynamic$dynEGA$population)
```



Do *variables* cluster into dimensions? Do we find the Big Five?

- Openness to Experience (community 6): partially replicated (O1, O3-O8)
- Conscientiousness (community 2): partially replicated (C1-C8, C11, C12)
- Extraversion (community 5): partially replicated (E1-E4, E9, E10, E12)
- Agreeableness (communities 4 and 7): split between two communities
- Neuroticism (community 1): perfectly replicated (N1–N12)
- Mixed (community 3): extraversion and openness to experience

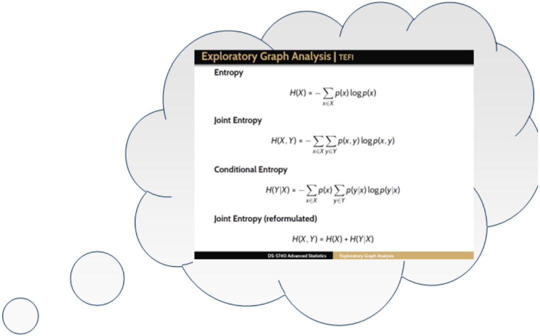
Quantifying Similarity of Communities

Normalized mutual information

$$NMI(C_{theo}, C_{est}) = \frac{2 \times I(C_{theo}, C_{est})}{[H(C_{theo}) + H(C_{est})]}$$

- entropy: $H(X) = -\sum_{x \in X} p(x) \log p(x)$
- mutual information: $I(X, Y) = H(X, Y) - H(X|Y) - H(Y|X)$

Dynamic Exploratory Graph Analysis



Exploratory Graph Analysis | TEFI

Entropy

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

Joint Entropy

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y)$$

Conditional Entropy

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log p(y|x)$$

Joint Entropy (reformulated)

$$H(X, Y) = H(X) + H(Y|X)$$

DS-5740 Advanced Statistics | Exploratory Graph Analysis

Dynamic Exploratory Graph Analysis

```r
# Set empirical memberships
empirical <- bfi2_dynamic$dynEGA$population$wc
names(empirical) <- gsub(".Ord1", "", names(empirical))

# Set theoretical memberships
theoretical <- empirical
theoretical[grep("O", names(theoretical))] <- 1
theoretical[grep("C", names(theoretical))] <- 2
theoretical[grep("E", names(theoretical))] <- 3
theoretical[grep("A", names(theoretical))] <- 4
theoretical[grep("N", names(theoretical))] <- 5

# NMI
igraph::compare(empirical, theoretical, method = "nmi")
```

```
[1] 0.755034
```

0 = independent community solutions

1 = perfect match

Is our value good?

```r
# Set empirical memberships
empirical <- bfi2_dynamic$dynEGA$population$wc
names(empirical) <- gsub(".Ord1", "", names(empirical))

# Set theoretical memberships
theoretical <- empirical
theoretical[grep("O", names(theoretical))] <- 1
theoretical[grep("C", names(theoretical))] <- 2
theoretical[grep("E", names(theoretical))] <- 3
theoretical[grep("A", names(theoretical))] <- 4
theoretical[grep("N", names(theoretical))] <- 5

# NMI
igraph::compare(empirical, theoretical, method = "nmi")
```

```
[1] 0.755034
```

0 = independent community solutions

1 = perfect match

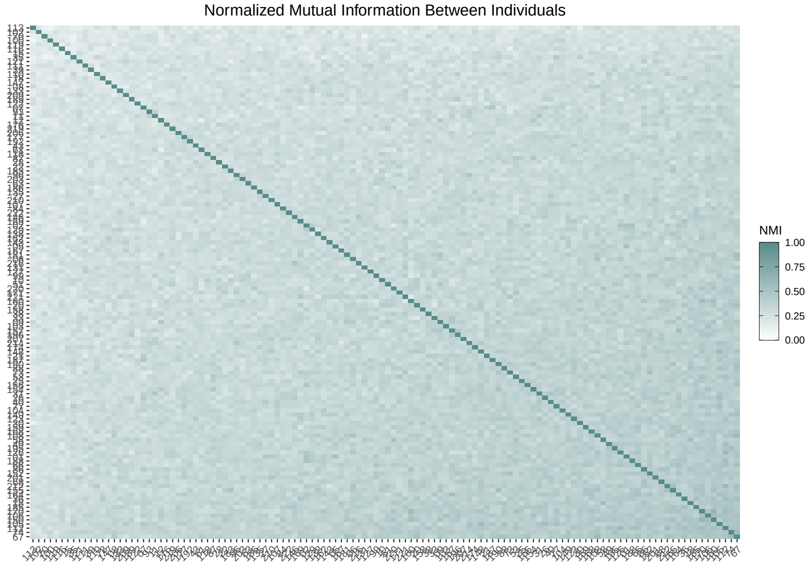
Is our value good? … it depends

Do *variables* cluster into the same dimensions for each person?

```r
# Summary for individuals
summary(bfi2_dynamic$dynEGA$individual)
```

```
Individual

Model: GLASSO (EBIC)
Correlations: auto
Unidimensional Method:  Louvain

----

Number of cases:  122

Median dimensions: 7

             5  6  7  8
Frequency:   6 54 43 19
```

Dynamic Exploratory Graph Analysis



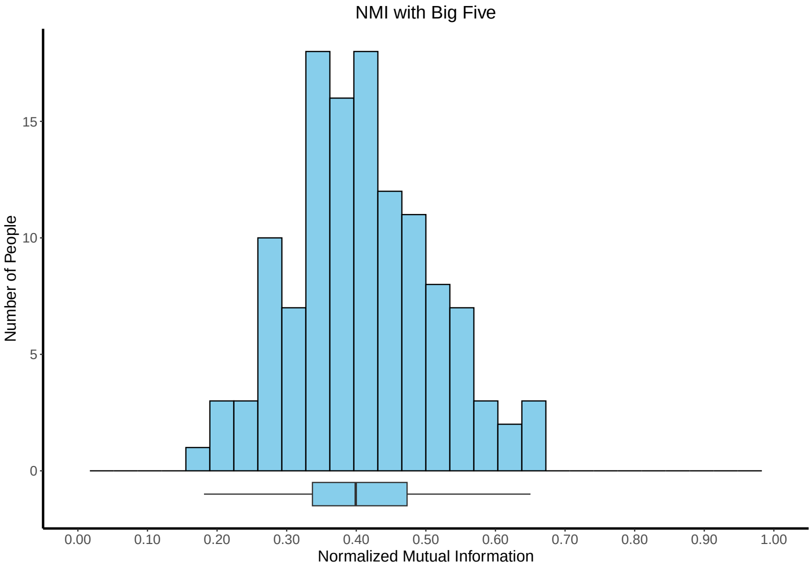
Do *variables* cluster into the same dimensions for each person?

NMI Descriptives

- mean = 0.323
- standard deviation = 0.071
- range = 0.093, 0.618

Doesn't seem like it…

What about each person and the Big Five?



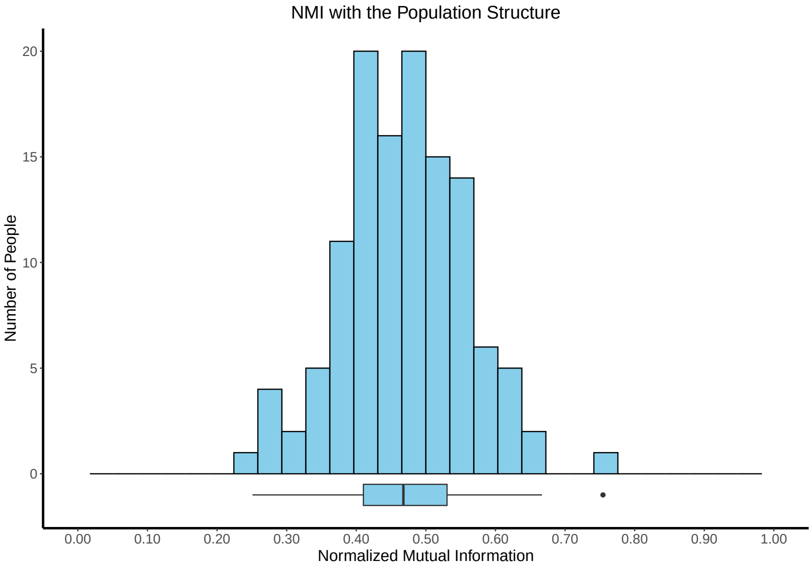
What about each person and the Big Five?

NMI Descriptives

- mean = 0.406
- standard deviation = 0.102
- range = 0.181, 0.65

Doesn’t seem like it…

What about each person and the population structure?



What about each person and the Big Five?

NMI Descriptives

- mean = 0.47
- standard deviation = 0.089
- range = 0.251, 0.754

Maybe one person? But not really...

Do *people* cluster into sub-groups or *types*?

- Can people be grouped based on similar network (not necessarily community) structures?
- Provides insights into *types* of people that might exist in our sample
- **Goal**: Identify meaningful groups that we can compare and potentially use as "natural" differences in an experiment
- May have implications for interventions or (clinical) treatments

Dynamic Exploratory Graph Analysis

Do *people* cluster into sub-groups or *types*?

- (Quantum) Jensen-Shannon Distance: computes distance or similarity between two network structures

- After hierarchical clustering can be applied to identify groups

(Quantum) Jensen-Shannon Distance

Starts with computing Von Neumann entropy of network

$$h_A = -\mathrm{Tr}[\mathcal{L}_G \log_2 \mathcal{L}_G]$$

- Tr = trace (sum of the diagonal)

- $\mathcal{L}_G$ = combinatorial Laplacian matrix: $c \times (D - A)$
  - A = network
  - D = sum of each variable's connection in the network on a diagonal matrix
  - $c = \frac{1}{\sum A}$

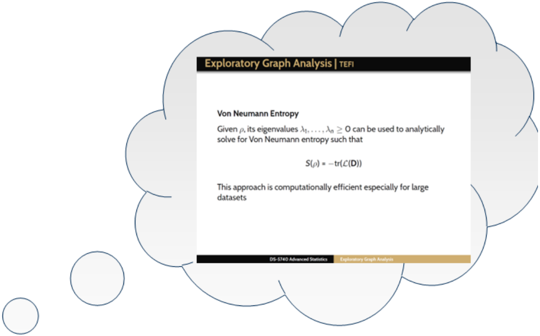
(Quantum) Jensen-Shannon Distance

Starts with computing Von Neumann entropy of network

$$h_A = -\sum_{i=1}^{N} \lambda_i \log_2(\lambda_i)$$

- λ_i = eigenvalues of $\mathcal{L}_G$

Dynamic Exploratory Graph Analysis



Exploratory Graph Analysis | TEFI

Von Neumann Entropy

Given ρ, its eigenvalues $\lambda_1, \ldots, \lambda_n \geq 0$ can be used to analytically solve for Von Neumann entropy such that

$$S(\rho) = -\mathrm{tr}(\mathcal{L}(\mathbf{D}))$$

This approach is computationally efficient especially for large datasets

DS-5740 Advanced Statistics | Exploratory Graph Analysis

(Quantum) Jensen-Shannon Distance

Starts with computing Von Neumann entropy of network

$$\mathcal{D}_{JS}(\rho||\sigma) = h(\mu) - \frac{1}{2}[h(\rho) + h(\sigma)]$$

- h = Von Neumann entropy of combinatorial Laplacian matrix
- μ = average combinatorial Laplacian matrix of network ρ and σ
- $\sqrt{\mathcal{D}_{JS}(\rho||\sigma)}$ = (Quantum) Jensen-Shannon Distance
  - Bounded between 0 and 1

Hierarchical Clustering

1. Uses agglomerative or “bottom-up” method on the Jensen-Shannon Distance

2. Applies the complete linkage function

$$max_{i,j}\, d(\mathbf{X_i}, \mathbf{Y_j})$$

3. Join observations/clusters that are most similar of all possible distance values (i.e., lowest value)

4. Repeat 2. and 3. until there is one cluster

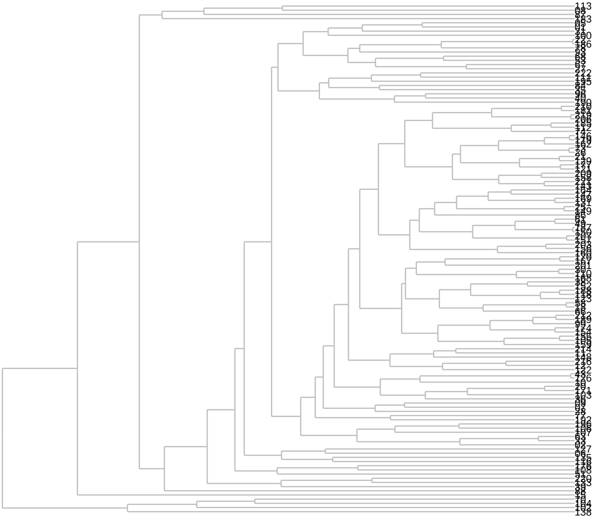
Hierarchical Clustering

- Through this process, a dendrogram or tree-like structure is created with “roots” and “branches”
- A “cut” can be made on these branches to obtain the clusters (from 1 to $n-1$)
- A criterion measure is computed for each cut and the cut that has the best criterion is selected
- In the present application, modularity is used

Information Theory Clustering

```r
# Compute clusters
bfi2_clusters <- infoCluster(bfi2_dynamic)

# Summary
summary(bfi2_clusters)
```

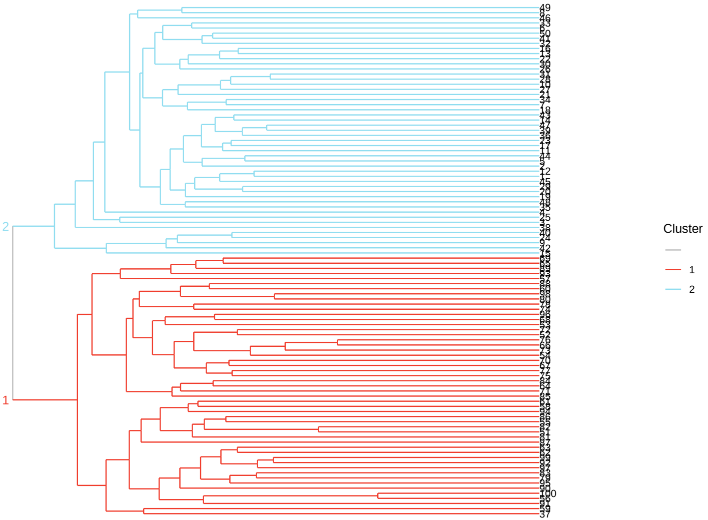
Dynamic Exploratory Graph Analysis



not our results – example when there *is* multiple clusters

```
Number of cases:  122
Number of clusters:  122

 01  02  05  06  07  08  09  10 100 102 103 104 105 106 107
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
108  11 110 111 112 113 116 118 119  12 121 122 123 126 127
 16  17  18  19  20  21  22  23  24  25  26  27  28  29  30
129 131 133 135 138 143 146 147 148 149  15 150 152 154 155
 31  32  33  34  35  36  37  38  39  40  41  42  43  44  45
156 157 158 159 160 162 164 167 168 169 170 171 174 176 177
 46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
 18 181 183 185 186 187 188 189 190 192 195 196  20 201 203
 61  62  63  64  65  66  67  68  69  70  71  72  73  74  75
206 207 209  21 210 212 214 216 219  22 220 221 222  25  26
 76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
 27  28  30  31  33  35  36  38  39  40  43  49  51  53  58
 91  92  93  94  95  96  97  98  99 100 101 102 103 104 105
 61  63  66  67  68  70  73  74  77  78  84  86  87  88  93
106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
 96  99
121 122
```

Dynamic Exploratory Graph Analysis



On Single Clusters

Our found a **single** cluster based on modularity

Single clusters are tricky because if all values are *relatively* equidistant then a single cluster will be returned

However, if all clusters are *relatively* equidistant, then it's also possible that the clustering is **random**

Therefore, we need a statistical test against random to determine whether we have a single cluster or no clusters

Single or Random Cluster Approach

1. Generate random networks by shuffling edges randomly in each individual's network such that the same *number* of edges exist but they are in a *different* arrangement
2. Compute JSD between each individual's random network
3. Compute a paired samples *t*-test using the paired values of actual JSD and random JSD

4. Interpret the test (actual - random; `$single.cluster.test$t.test`):

a. Positive values: the distances between the actual networks are **greater than** the random networks suggesting **no clusters**

b. Negative values: the distances between the actual networks are **less than** the random networks suggesting a **single cluster**

c. $p < 0.05$ should be true *and* $p_{adaptive} < 0.05$ should *also* be true

Cohen's *d*

- small (0.20)
- moderate (0.50)
- large (0.80)

Takeaways

We didn't find any clusters!

This result suggests that each person in our sample is *unique*

What implications does that hold for measurement?