# Meaning in colexification: beyond single edges and towards a network perspective

Loading libraries

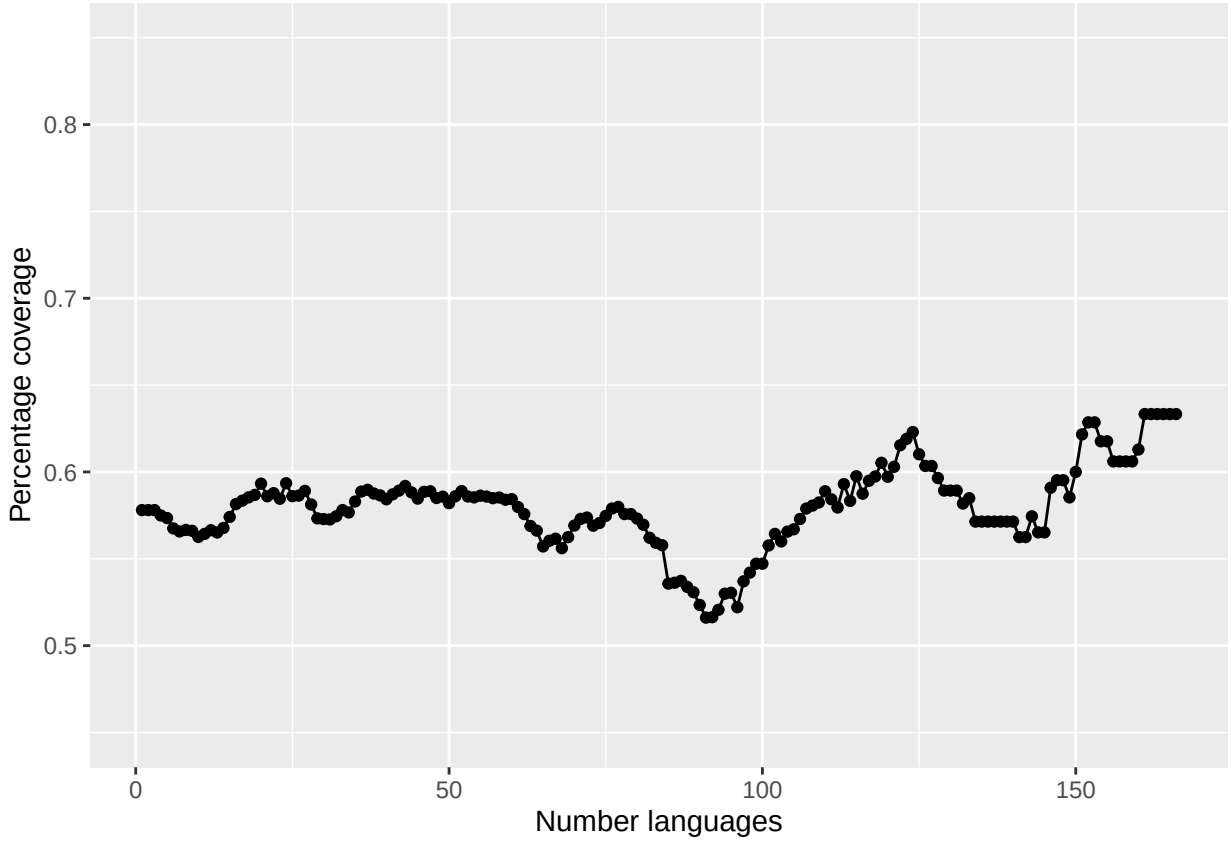
Loading data

Loading function for computing the distances in the network

Computing the distances in the network (it might take some time)

Loading and preprocessing the distances on the network

Figure 2 SI: Coverage of MTurk questions on Clics3



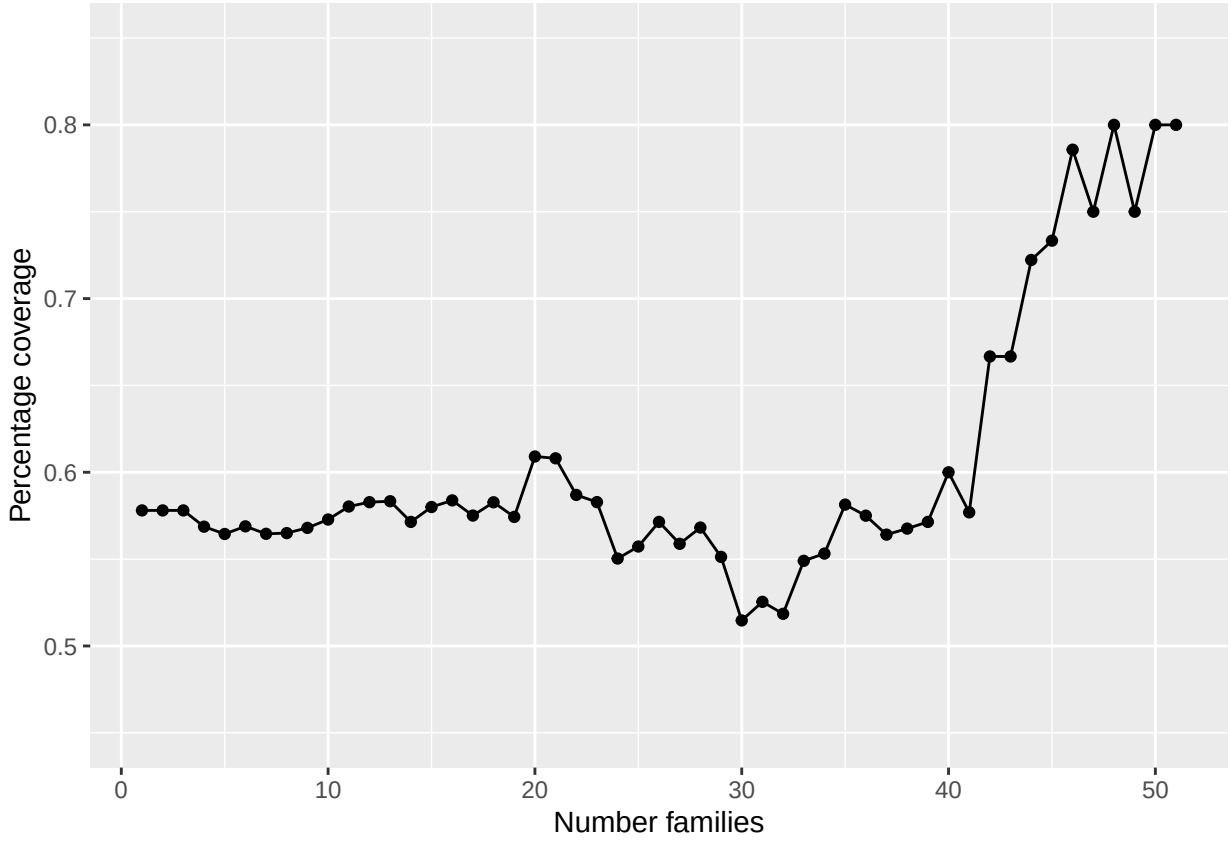



Table 1 - correlation of MTurk results with ground truth datasets (SimLex-999, SimVerb-3500, MEN)

| | dataset | corr_mean | corr_ci1 | corr_ci2 | n | pval |
|---|---|---|---|---|---|---|
| cor | SimLex-999 | 0.50 | 0.27 | 0.68 | 53 | <0.001 |
| cor1 | SimVerb-3500 | 0.57 | 0.45 | 0.67 | 145 | <0.001 |
| cor2 | MEN | 0.63 | 0.44 | 0.77 | 56 | <0.001 |

Table 1 si - correlation of the word association data with MTurk annotations

| | dataset | corr_mean | corr_ci1 | corr_ci2 | n | pval |
|---|---|---|---|---|---|---|
| cor | SWOW | 0.26 | 0.20 | 0.31 | 1024 | <0.001 |
| cor1 | USF | 0.25 | 0.19 | 0.31 | 880 | <0.001 |

LINK LEVEL

Table 2 - correlation of similarity data with colexification strength (link level) Table 2 SI rows 1, 2, 3, 4, 7 (SimLex, SimVerb, MEN, FastText and MTurk annotations) second column - number of overlapping edges Table 3 SI: correlation with the mode of the MTurk annotations

| | dataset | corr_lang | corr_lang_ci1 | corr_lang_ci2 | pval_lang | corr_fam | corr_fam_ci1 | corr_fam_ci2 | pval_fam | n |
|---|---|---|---|---|---|---|---|---|---|---|
| cor | SimLex-999 | 0.27 | 0.01 | 0.49 | 0.04 | 0.38 | 0.14 | 0.58 | 0.003 | 59 |
| cor1 | SimVerb-3500 | 0.23 | 0.08 | 0.37 | 0.003 | 0.25 | 0.10 | 0.38 | 0.001 | 168 |

| | dataset | corr_lang | corr_lang_ci1 | corr_lang_ci2 | pval_lang | corr_fam | corr_fam_ci1 | corr_fam_ci2 | pval_fam | n |
|---|---|---|---|---|---|---|---|---|---|---|
| cor2 | MEN | 0.32 | 0.06 | 0.54 | 0.016 | 0.35 | 0.09 | 0.56 | 0.009 | 56 |
| cor3 | FastText | 0.16 | 0.12 | 0.20 | <0.001 | 0.19 | 0.15 | 0.23 | <0.001 | 2441 |
| cor4 | Annotations | 0.25 | 0.21 | 0.29 | <0.001 | 0.32 | 0.28 | 0.35 | <0.001 | 2441 |

| dataset | n_link | perc_link |
|---|---|---|
| SimLex-999 | 59 | 1.4 |
| SimVerb-3500 | 168 | 4.0 |
| MEN | 56 | 1.3 |
| FastText | 2441 | 57.7 |
| Annotations | 2441 | 57.7 |

| | corr_lang | corr_lang_ci1 | corr_lang_ci2 | pval_lang | corr_fam | corr_fam_ci1 | corr_fam_ci2 | pval_fam | n |
|---|---|---|---|---|---|---|---|---|---|
| cor | 0.22 | 0.18 | 0.25 | <0.001 | 0.27 | 0.23 | 0.31 | <0.001 | 2441 |

Table 3 - correlation of colexification strength with word association tasks Table 2 si rows 5,6 (SWOW and USF), second column (edges)

| | dataset | corr_lang | corr_lang_ci1 | corr_lang_ci2 | pval_lang | corr_fam | corr_fam_ci1 | corr_fam_ci2 | pval_fam | n |
|---|---|---|---|---|---|---|---|---|---|---|
| cor | SWOW | 0.15 | 0.09 | 0.20 | <0.001 | 0.18 | 0.13 | 0.24 | <0.001 | 1189 |
| cor1 | USF | 0.08 | 0.02 | 0.14 | 0.012 | 0.14 | 0.08 | 0.20 | <0.001 | 982 |

| dataset | n_link | perc_link |
|---|---|---|
| SWOW | 1189 | 28.1 |
| USF | 982 | 23.2 |

Table 4 SI - Results on a common dataset (FastText, SWOW, annotations)

| | dataset | corr_lang | corr_lang_ci1 | corr_lang_ci2 | pval_lang | corr_fam | corr_fam_ci1 | corr_fam_ci2 | pval_fam | dist_lang | dist_lang_ci1 | dist_lang_ci2 | pval_dist_lang | dist_fam | dist_fam_ci1 | dist_fam_ci2 | pval_dist_fam |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cor | Annotations | 0.05 | 0.17 | <0.001 | 0.14 | 0.08 | 0.20 | <0.001 | 0.12 | 0.06 | 0.18 | <0.001 | 0.11 | 0.05 | 0.17 | <0.001 | |
| cor1 | FastText | -0.04 | -0.02 | 0.11 | 0.15 | 60.07 | 0.01 | 0.13 | 0.02 | 50.12 | 0.06 | 0.18 | <0.001 | 0.14 | 0.08 | 0.20 | <0.001 |
| cor2 | SWOW | 0.14 | 0.08 | 0.20 | <0.001 | 0.18 | 0.12 | 0.24 | <0.001 | 0.20 | 0.14 | 0.26 | <0.001 | 0.20 | 0.14 | 0.25 | <0.001 |

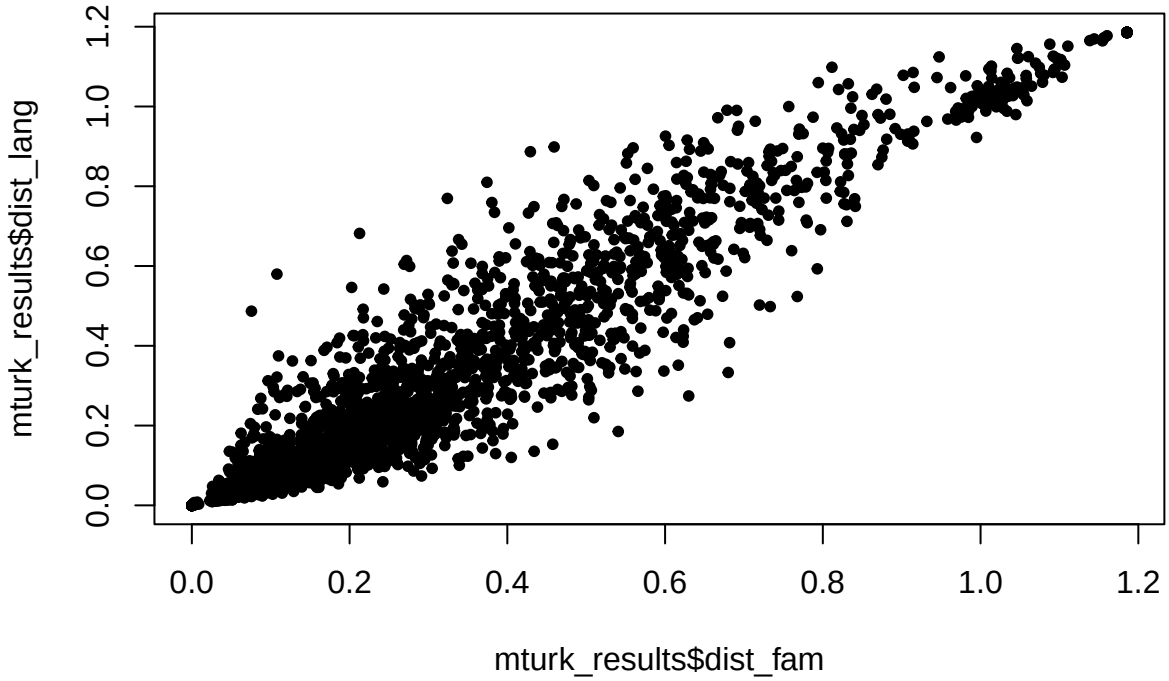
NETWORK LEVEL

Figure 3- correlation between distances and colexification weigths

```
##
##  Pearson's product-moment correlation
##
## data:  mturk_results$dist_fam and mturk_results$dist_lang
## t = 162.74, df = 2658, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  0.9497103 0.9566520
## sample estimates:
##       cor
## 0.9533069
```

```
##
##  Pearson's product-moment correlation
##
## data:  mturk_results$FamilyWeight and mturk_results$LanguageWeight
## t = 78.527, df = 2439, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8348734 0.8573877
## sample estimates:
##       cor
## 0.8465086
```



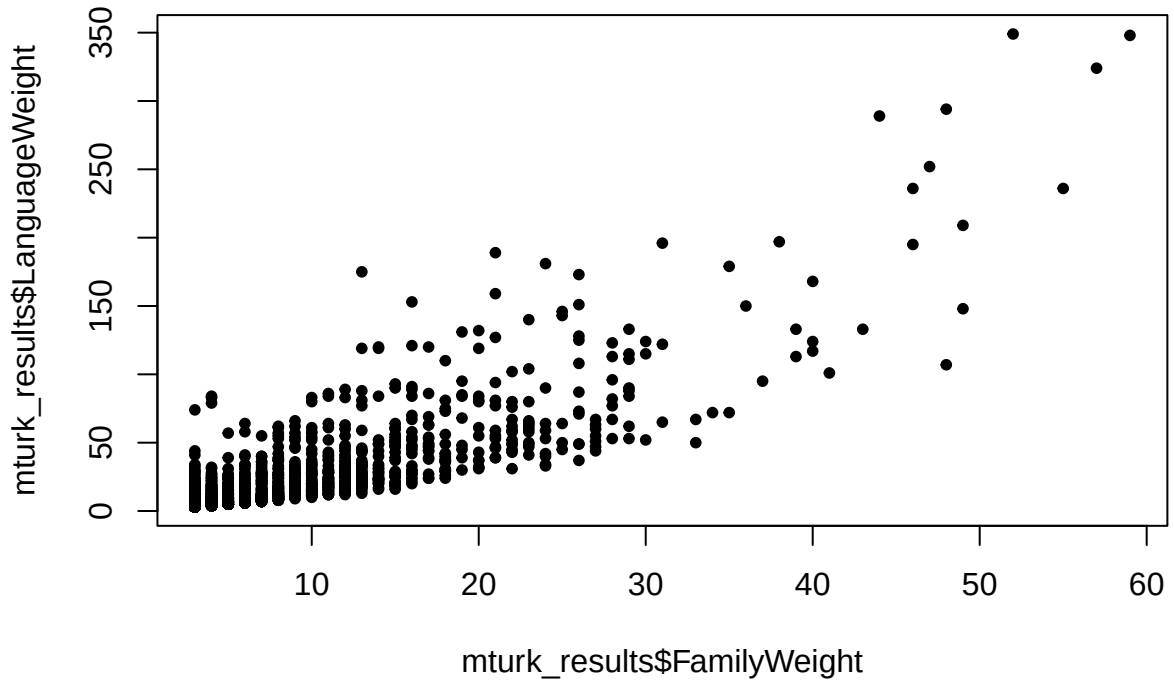



Table 4 - correlation of similarity data with distance on language weights (network level) Table 2 SI rows 1, 2, 3, 4, 7 (SimLex, SimVerb, MEN, FastText and MTurk annotations) third column - number of overlapping distances and Table 6 SI: correlation with the distance computed with family weights

```
##
##   Results of a comparison of two overlapping correlations based on dependent groups
##
## Comparison between r.jk (mean, dist_lang) = 0.3129 and r.jh (mean, LanguageWeight) = 0.2519
## Difference: r.jk - r.jh = 0.061
## Related correlation: r.kh = 0.4237
## Data: mturk_results: j = mean, k = dist_lang, h = LanguageWeight
## Group size: n = 2441
## Null hypothesis: r.jk is equal to r.jh
## Alternative hypothesis: r.jk is greater than r.jh (one-sided)
## Alpha: 0.05
##
## pearson1898: Pearson and Filon's z (1898)
##   z = 2.9648, p-value = 0.0015
##   Null hypothesis rejected
##
## hotelling1940: Hotelling's t (1940)
##   t = 2.9809, df = 2438, p-value = 0.0015
##   Null hypothesis rejected
##
## williams1959: Williams' t (1959)
##   t = 2.9654, df = 2438, p-value = 0.0015
##   Null hypothesis rejected
##
## olkin1967: Olkin's z (1967)
##   z = 2.9648, p-value = 0.0015
##   Null hypothesis rejected
##
## dunn1969: Dunn and Clark's z (1969)
```

```
##   z = 2.9625, p-value = 0.0015
##   Null hypothesis rejected
##
## hendrickson1970: Hendrickson, Stanley, and Hills' (1970) modification of Williams' t (1959)
##   t = 2.9809, df = 2438, p-value = 0.0015
##   Null hypothesis rejected
##
## steiger1980: Steiger's (1980) modification of Dunn and Clark's z (1969) using average correlations
##   z = 2.9614, p-value = 0.0015
##   Null hypothesis rejected
##
## meng1992: Meng, Rosenthal, and Rubin's z (1992)
##   z = 2.9604, p-value = 0.0015
##   Null hypothesis rejected
##   95% confidence interval for r.jk - r.jh: 0.0224 0.1102
##   Null hypothesis rejected (Lower boundary > 0)
##
## hittner2003: Hittner, May, and Silver's (2003) modification of Dunn and Clark's z (1969) using a bac
##   z = 2.9612, p-value = 0.0015
##   Null hypothesis rejected
##
## zou2007: Zou's (2007) confidence interval
##   95% confidence interval for r.jk - r.jh: 0.0206 0.1013
##   Null hypothesis rejected (Lower boundary > 0)
```

| | dataset | corr_dist_lang | corr_dist_lang_ci1 | corr_dist_lang_ci2 | pval_dist_lang | n |
|---|---|---|---|---|---|---|
| cor | SimLex-999 | 0.47 | 0.36 | 0.56 | <0.001 | 220 |
| cor1 | SimVerb-3500 | 0.49 | 0.42 | 0.55 | <0.001 | 525 |
| cor2 | MEN | 0.40 | 0.31 | 0.48 | <0.001 | 382 |
| cor3 | FastText | 0.30 | 0.26 | 0.33 | <0.001 | 2641 |
| cor4 | Annotations | 0.37 | 0.34 | 0.40 | <0.001 | 2660 |

| dataset | n_dist | perc_dist |
|---|---|---|
| SimLex-999 | 220 | 0.0 |
| SimVerb-3500 | 525 | 0.0 |
| MEN | 382 | 0.0 |
| FastText | 2641 | 0.2 |
| Annotations | 2660 | 0.2 |

| | dataset | corr_dist_fam | corr_dist_fam_ci1 | corr_dist_fam_ci2 | pval_dist_fam | n |
|---|---|---|---|---|---|---|
| cor | SimLex-999 | 0.46 | 0.35 | 0.56 | <0.001 | 220 |
| cor1 | SimVerb-3500 | 0.49 | 0.42 | 0.55 | <0.001 | 525 |
| cor2 | MEN | 0.41 | 0.32 | 0.49 | <0.001 | 382 |
| cor3 | FastText | 0.31 | 0.28 | 0.34 | <0.001 | 2641 |
| cor4 | Annotations | 0.37 | 0.34 | 0.41 | <0.001 | 2660 |

Table 2 SI rows 5,6 (USF, SWOW) third column - number of overlapping distances

| dataset | n_dist | perc_dist |
|---|---|---|
| SWOW | 10183 | 0.9 |
| USF | 12505 | 1.1 |

Table 5 - correlation of word association tasks and distance on language weights Table 7 SI - correlation of association tasks data and distance on family weights

| | dataset | corr_dist_lang | corr_dist_lang_ci1 | corr_dist_lang_ci2 | pval_dist_lang | n |
|---|---|---|---|---|---|---|
| cor | SWOW | 0.29 | 0.27 | 0.30 | <0.001 | 10183 |
| cor1 | USF | 0.19 | 0.17 | 0.21 | <0.001 | 12505 |

| | dataset | corr_dist_fam | corr_dist_fam_ci1 | corr_dist_fam_ci2 | pval_dist_fam | n |
|---|---|---|---|---|---|---|
| cor | SWOW | 0.29 | 0.27 | 0.3 | <0.001 | 10183 |
| cor1 | USF | 0.19 | 0.17 | 0.2 | <0.001 | 12505 |

Table 6 - linear regression models for distance on the language weights (network level)

```
##
## Call:
## lm(formula = mturk_results$mean ~ scale(mturk_results$cossim) +
##     scale(mturk_results$dist_lang))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.86029 -0.60103  0.00191  0.60361  2.13918
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     3.04507    0.01640  185.63   <2e-16 ***
## scale(mturk_results$cossim)     0.36750    0.01704   21.56   <2e-16 ***
## scale(mturk_results$dist_lang)  0.23172    0.01704   13.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8286 on 2553 degrees of freedom
##   (104 observations deleted due to missingness)
## Multiple R-squared:  0.2554, Adjusted R-squared:  0.2548
## F-statistic: 437.8 on 2 and 2553 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = mturk_results$mean ~ scale(mturk_results$cossim))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.70795 -0.63517  0.00672  0.64982  2.47019
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.05426    0.01697  179.98   <2e-16 ***
```

```
## scale(mturk_results$cossim)  0.43082    0.01697   25.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.858 on 2554 degrees of freedom
##   (104 observations deleted due to missingness)
## Multiple R-squared:  0.2014, Adjusted R-squared:  0.2011
## F-statistic: 644.3 on 1 and 2554 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = mturk_results$mean ~ scale(mturk_results$dist_lang))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.82891 -0.68647  0.02153  0.67464  2.27810
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.01244    0.01749  172.20   <2e-16 ***
## scale(mturk_results$dist_lang)  0.36004    0.01750   20.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9022 on 2658 degrees of freedom
## Multiple R-squared:  0.1374, Adjusted R-squared:  0.1371
## F-statistic: 423.4 on 1 and 2658 DF,  p-value: < 2.2e-16
```

| #Df | LogLik | Df | Chisq | Pr(>Chisq) |
|---|---|---|---|---|
| 4 | -3144.854 | NA | NA | NA |
| 3 | -3234.240 | -1 | 178.7726 | 0 |

Table 8 SI - linear regression models for distance on the family weights (network level)

```
##
## Call:
## lm(formula = mturk_results$mean ~ scale(mturk_results$cossim) +
##     scale(mturk_results$dist_fam))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.83256 -0.60600  0.00525  0.60277  2.13105
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.04416    0.01643  185.24   <2e-16 ***
## scale(mturk_results$cossim)    0.36609    0.01713   21.37   <2e-16 ***
## scale(mturk_results$dist_fam)  0.22847    0.01721   13.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.83 on 2553 degrees of freedom
##   (104 observations deleted due to missingness)
## Multiple R-squared:  0.253,  Adjusted R-squared:  0.2524
```

```
## F-statistic: 432.4 on 2 and 2553 DF,  p-value: < 2.2e-16
```

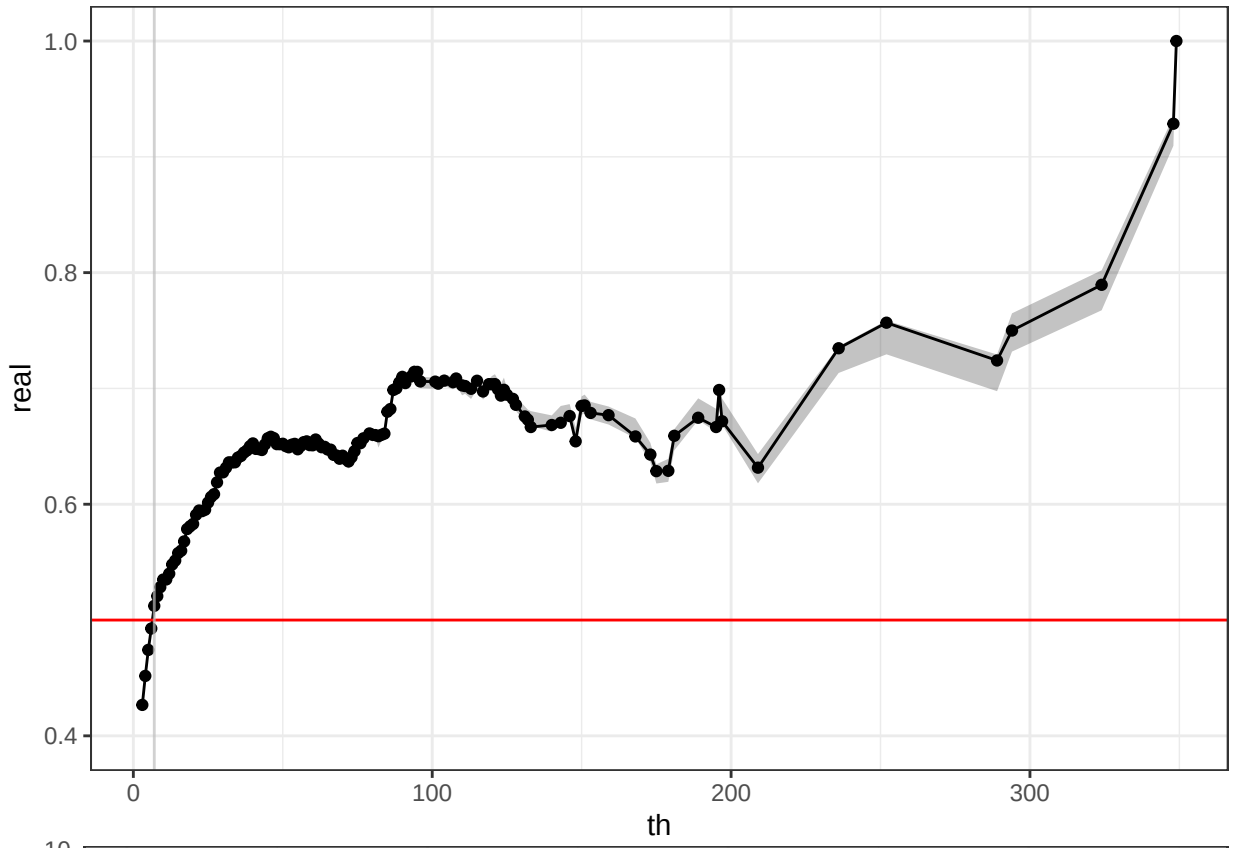
```
##
## Call:
## lm(formula = mturk_results$mean ~ scale(mturk_results$cossim))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.70795 -0.63517  0.00672  0.64982  2.47019
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.05426    0.01697  179.98   <2e-16 ***
## scale(mturk_results$cossim)  0.43082    0.01697   25.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.858 on 2554 degrees of freedom
##   (104 observations deleted due to missingness)
## Multiple R-squared:  0.2014, Adjusted R-squared:  0.2011
## F-statistic: 644.3 on 1 and 2554 DF,  p-value: < 2.2e-16
```

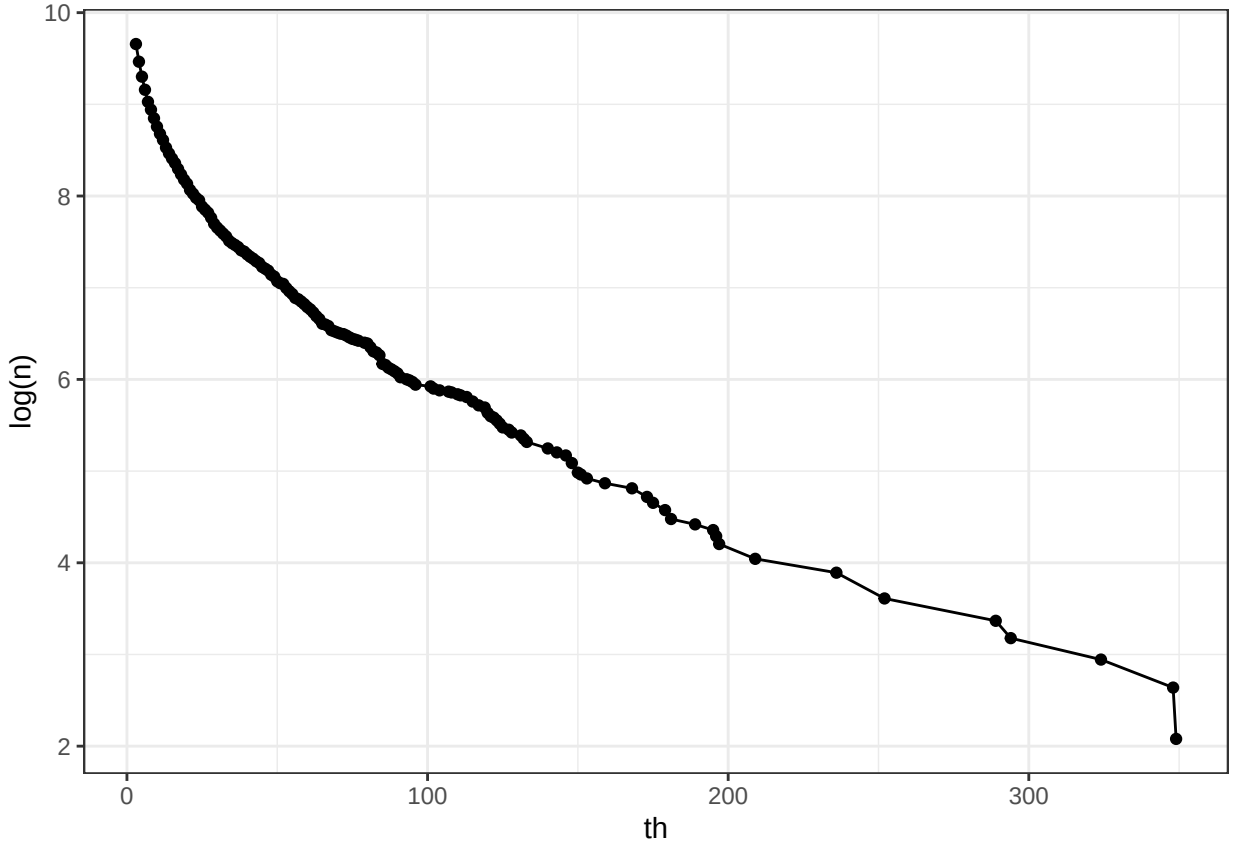
```
##
## Call:
## lm(formula = mturk_results$mean ~ scale(mturk_results$dist_fam))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89440 -0.68497  0.00716  0.67574  2.29344
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.01244    0.01747  172.40   <2e-16 ***
## scale(mturk_results$dist_fam)  0.36263    0.01748   20.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9012 on 2658 degrees of freedom
## Multiple R-squared:  0.1394, Adjusted R-squared:  0.1391
## F-statistic: 430.5 on 1 and 2658 DF,  p-value: < 2.2e-16
```

| #Df | LogLik | Df | Chisq | Pr(>Chisq) |
|---|---|---|---|---|
| 4 | -3148.924 | NA | NA | NA |
| 3 | -3234.240 | -1 | 170.6308 | 0 |

THRESHOLD FOR NOISE

Figure 4 - Estimation of the threshold for noise


1.0
0.8
0.6
0.4
real
0
100
200
300
th


10
8
6
4
2
log(n)
0
100
200
300
th

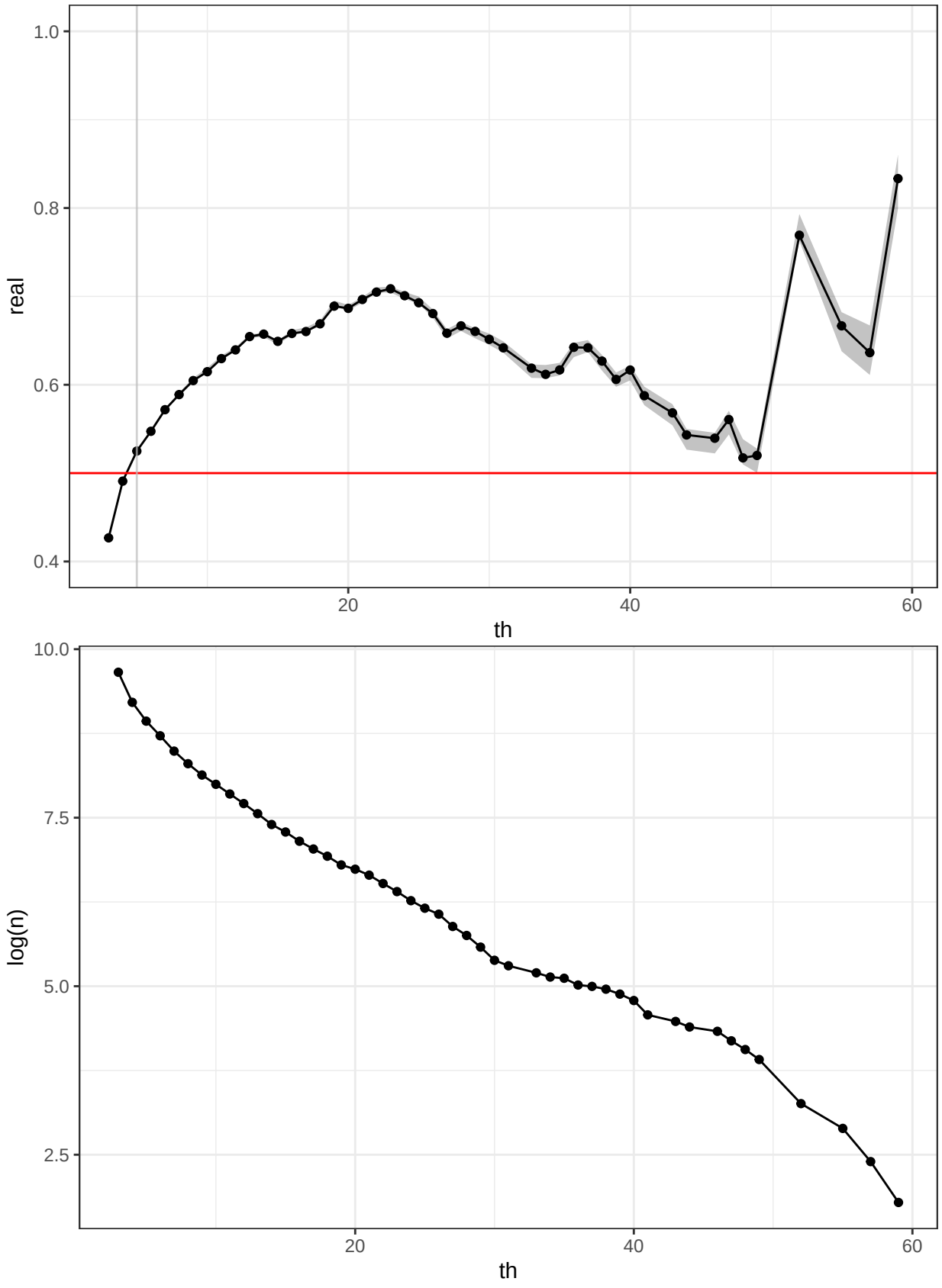



Figure 3 SI - estimation of the threshold for noise in the case of a relaxation of the definition of similarity

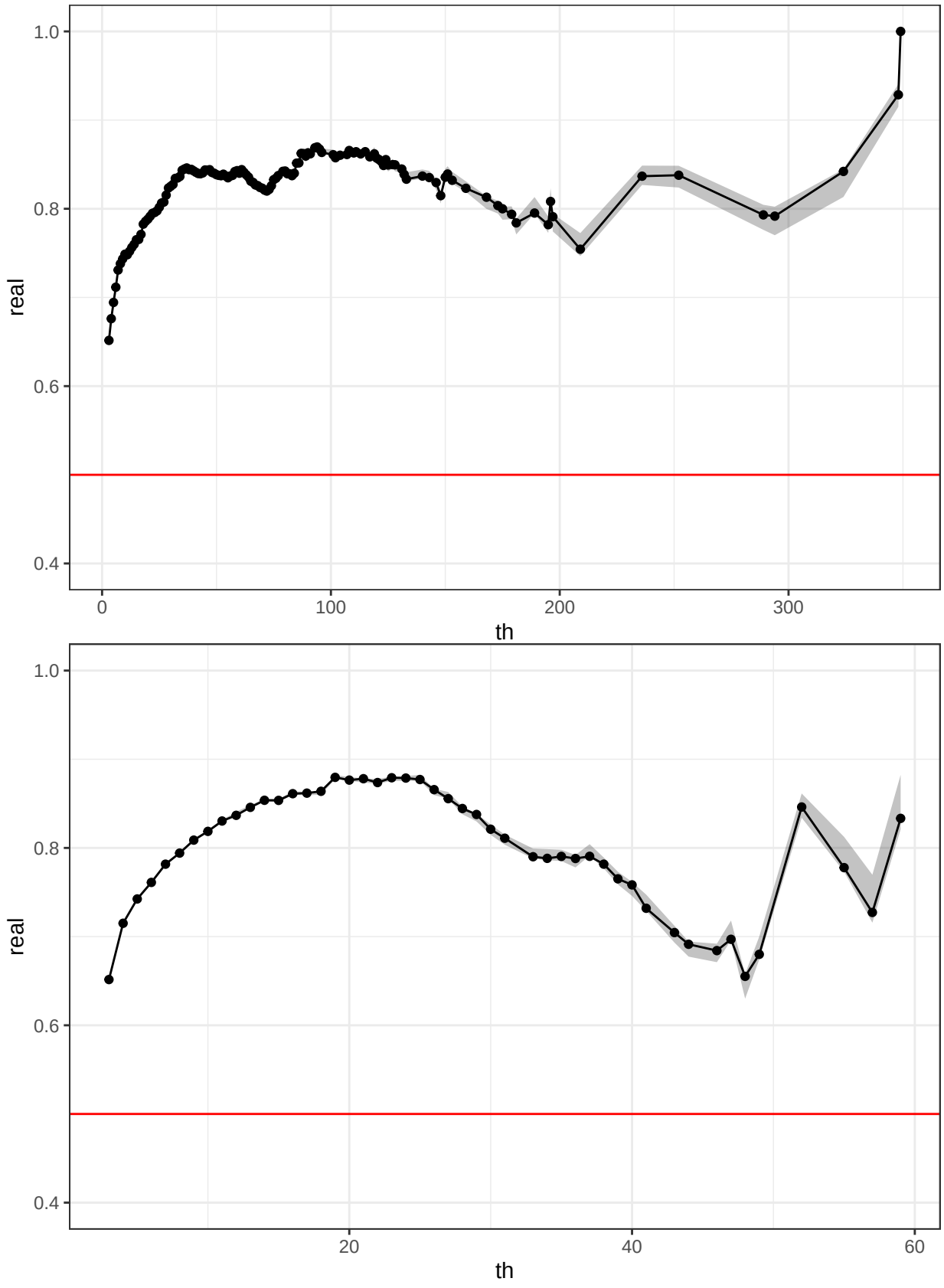



Figure 5 - heatmaps of the MTurk annotations

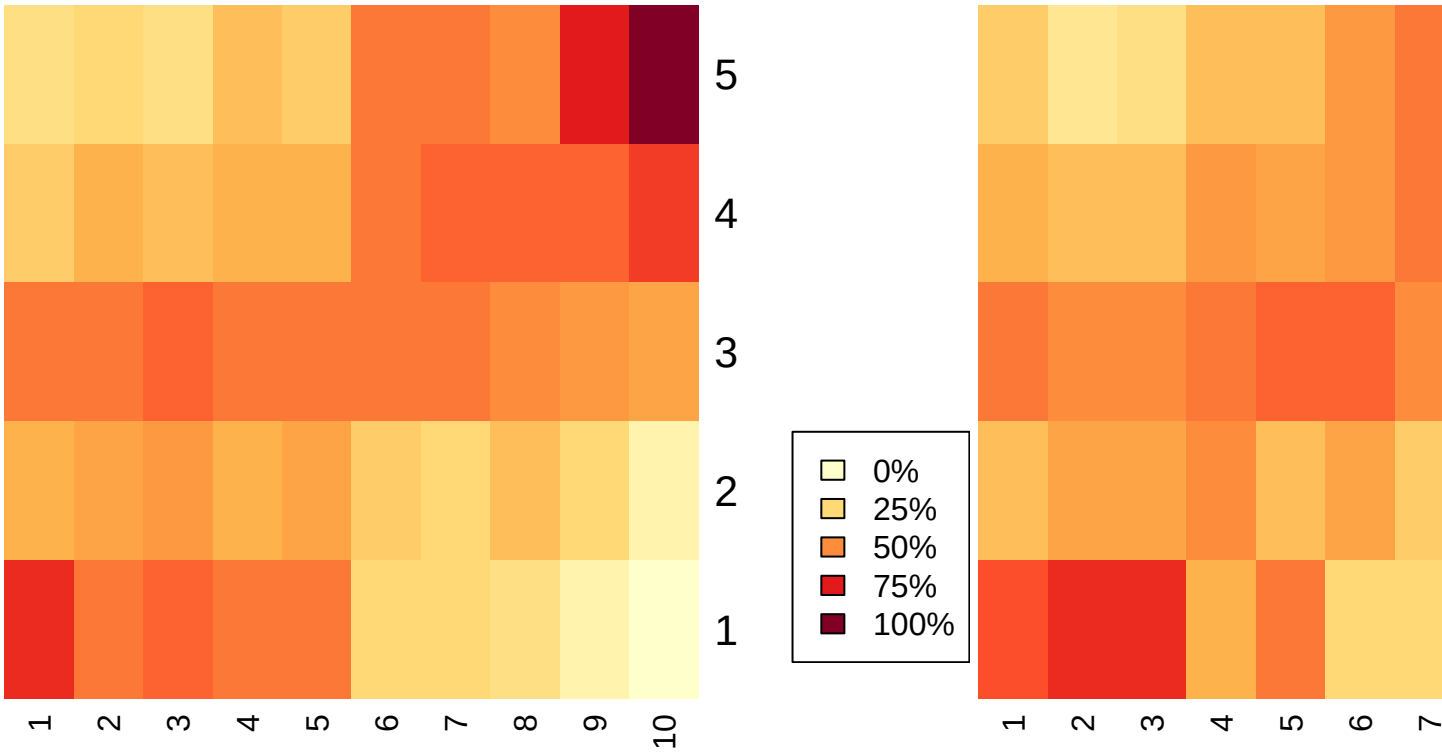



Figure 4 SI - heatmaps with distances