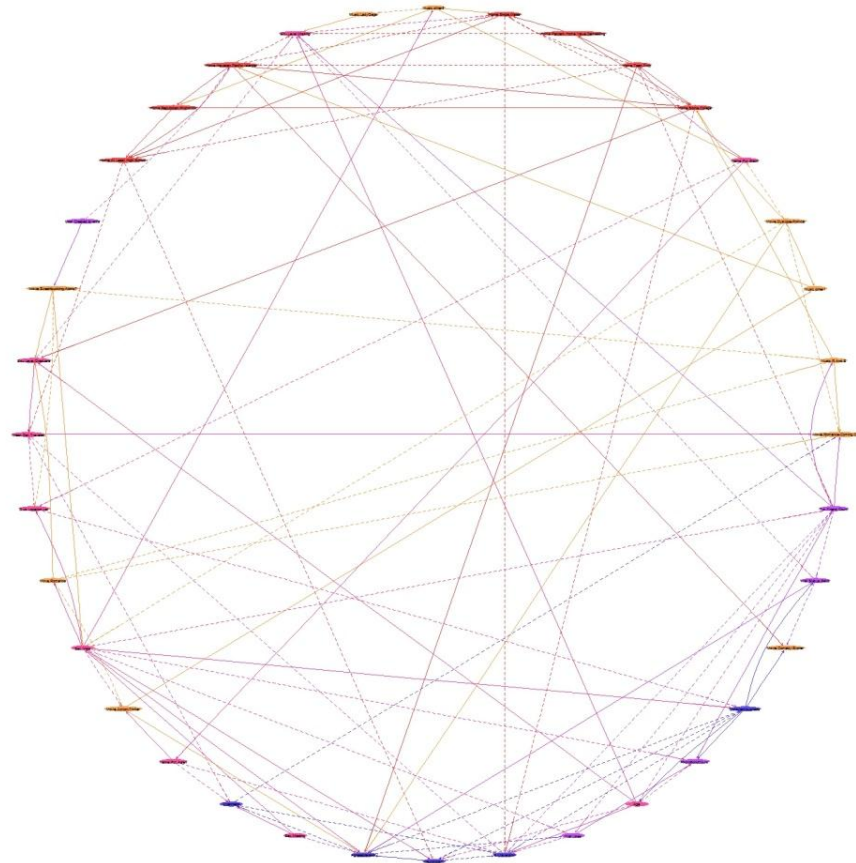


Leveraging online social network data and external data sources to predict personality

Daniel Chapsky



Overview

- ▶ Personality can be expressed through actions online, especially on online social networks (OSNs) like Facebook.
- ▶ Data on OSNs can be connected to external data sources for further inference.
- ▶ Machine learning to connect networks of information to predict personality
- ▶ Framework which can learn about the inferences as well as develop predictions



Overview - Steps

- ▶ Collect personality data on a sample
- ▶ Collect Facebook data of the sample
- ▶ Mash up Facebook data with external APIs to infer attributes, behaviours and culture of sample
- ▶ Generate a machine learned model which predicts personality through inferences



Tools

- ▶ An online personality quiz
- ▶ Collected Facebook data
- ▶ Online Data Sources
- ▶ Revolution R Enterprise version 4.0 (for academics)



R

- ▶ All steps besides quiz done in R
- ▶ Database connectivity
 - ▶ RMySQL,
- ▶ Web scraping / API connection
 - ▶ RCurl, RJSONIO, XML
- ▶ Inference through mashups
 - ▶ psych, geosphere



R Continued

- ▶ **Data Cleaning**
 - ▶ plyr, reshape2, bayestree, mice, tm, mvoutlier
- ▶ **Bayesian Network construction**
 - ▶ bnlearn, pcalg
- ▶ **Parallelization of optimization**
 - ▶ foreach, snow
- ▶ **Graphics**
 - ▶ Latticist, bnlearn, ggplot2



Personality

- ▶ Personality is the collection of behavioral and mental attributes that characterize an individual
- ▶ Actions and perspectives are considered expressions of underlying personality
- ▶ A portrait of an individual's personality can be captured by combining his/her
 - ▶ Attributes
 - ▶ Culture
 - ▶ Behaviors



Personality - Big 5 Theory

- ▶ Personality can be largely described by 5 personality factors
- ▶ The five factors compose a 5 dimensional personality space.
- ▶ Knowledge of these factors can be used to predict the attributes, views and behaviors of an individual



Personality – 5 Factors

- ▶ **Neuroticism**
 - ▶ Anxiety, Impulsivity
- ▶ **Extraversion**
 - ▶ Energy, stimulation through company of others
- ▶ **Openness (to experience)**
 - ▶ Academic curiosity, highly correlated with liberal political leaning
- ▶ **Agreeableness**
 - ▶ Compassion, desire for social cohesiveness
- ▶ **Conscientiousness**
 - ▶ Discipline, organization, motivation



Personality - Quiz Page

Bayesian Personality

Take the first quiz!

Tell your friends!

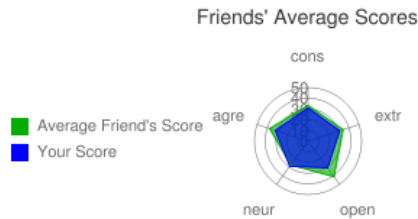
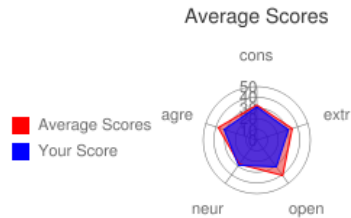
Take the second quiz!

Learn More

Second Quiz

First Quiz

Your last score
 Conscientiousness: 31
 Extraversion: 31
 Openness: 31
 Neuroticism: 29
 Agreeableness: 32



Dan Chapsky, Jeremy Koelmel and 20 others like this. Unlike · Admin Page



facebook

Search

Big Five Personality Test

The following pages contain phrases describing people's behaviors. Please use the rating scale next to each phrase to describe how accurately each statement describes you. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age.

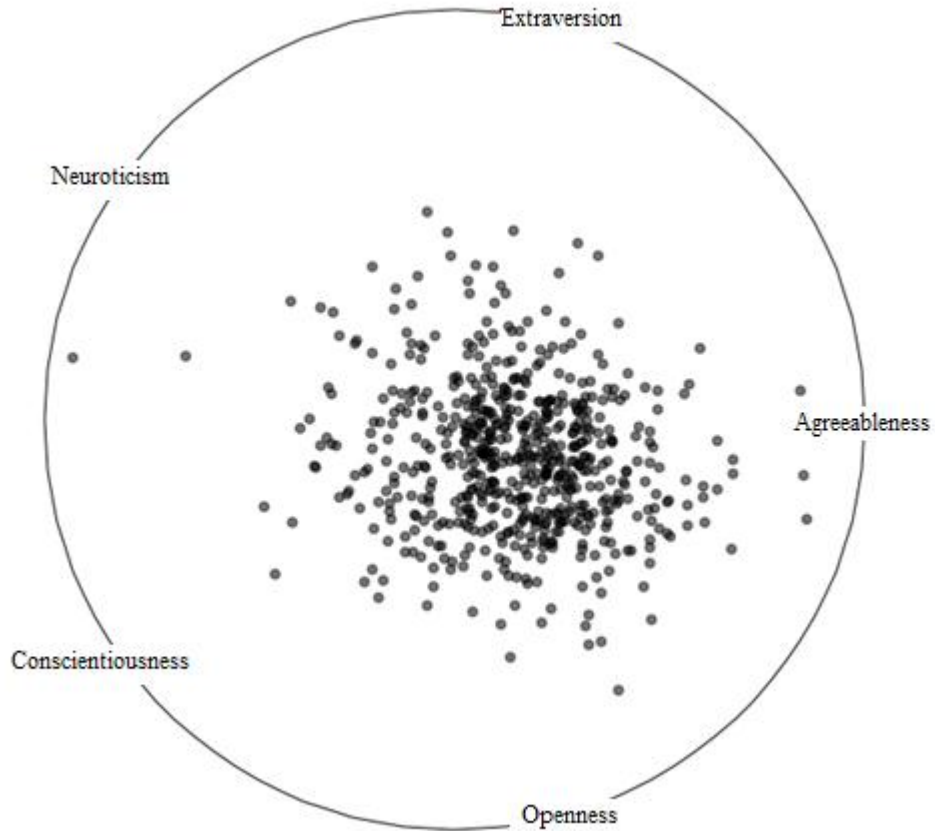
I...	Strongly Disagree		Neutral		Strongly Agree
Make friends easily.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Am always prepared.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Have difficulty understanding abstract ideas	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Pay attention to details.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Am not really interested in others.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Feel comfortable with myself.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Do just enough work to get by.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Carry the conversation to a higher level.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Rarely get irritated.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Love to read challenging material.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I...	Strongly Disagree		Neutral		Strongly Agree
Dislike myself.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Do not enjoy going to	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Personality – Quiz Results

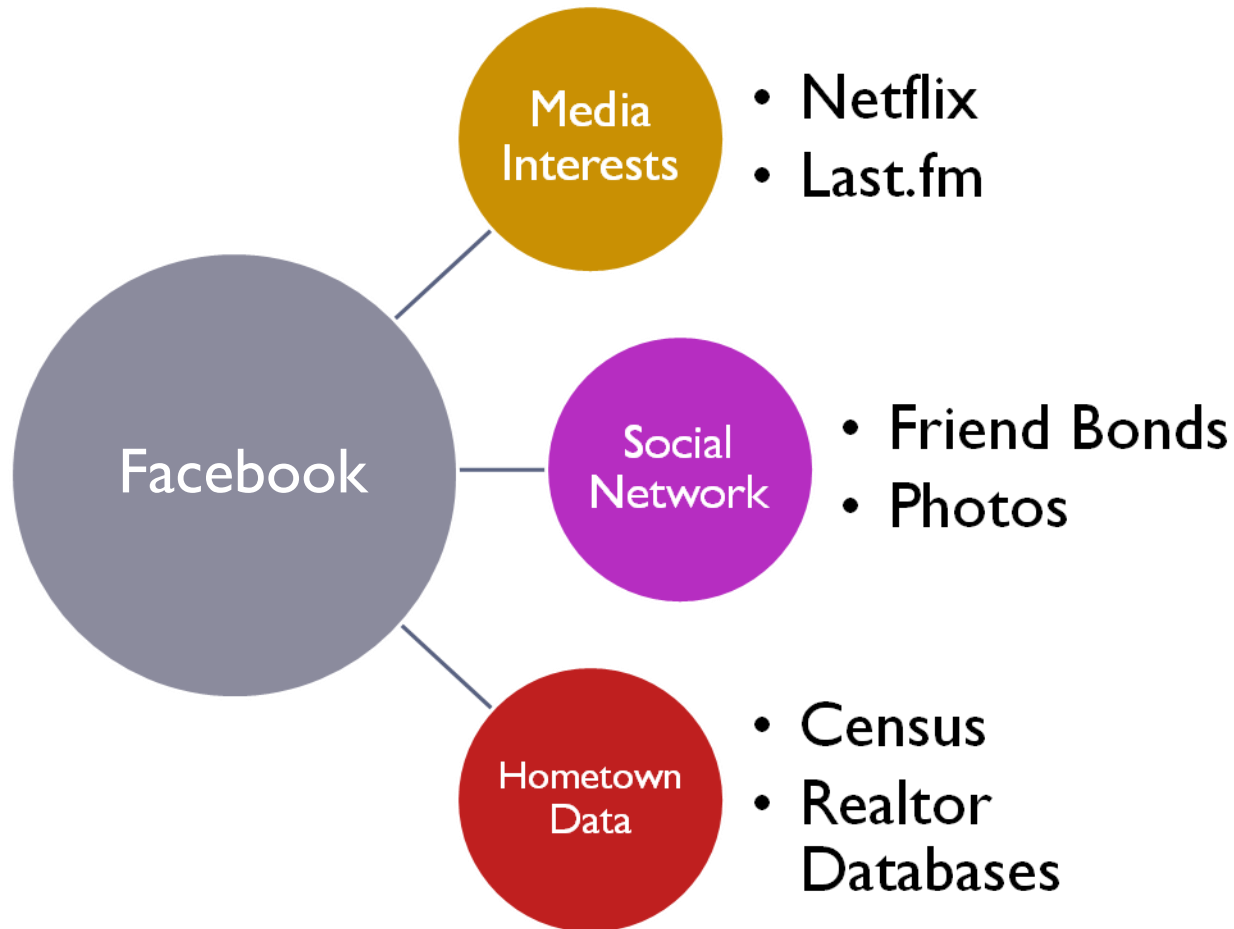
- ▶ 100 Question IPIP NEO-P-R
- ▶ Factor score 0- 50 scale
- ▶ 615 Respondents
- ▶ 35% Hampshire College students
- ▶ 35% Friends of Hampshire College students
- ▶ 30% recruited from online ads
- ▶ Skew towards politically liberal and middle class



Personality - Quiz Results



Inference



Inference - Media Preferences

Movies

- ▶ Netflix Data
- ▶ Preset list of possible genres
- ▶ Movies assigned Genres by Netflix
- ▶ 2100 movies analyzed

Music

- ▶ Last.fm data
- ▶ User generated tags
- ▶ Using tags voted on to be most popular
- ▶ 7500 bands analyzed

Libraries : RMySQL, RCurl, RJSONIO, XML



Inference - Media

Sample Music Factors

- ▶ Rock-80s
- ▶ Rap
- ▶ Metal
- ▶ Lady Gaga

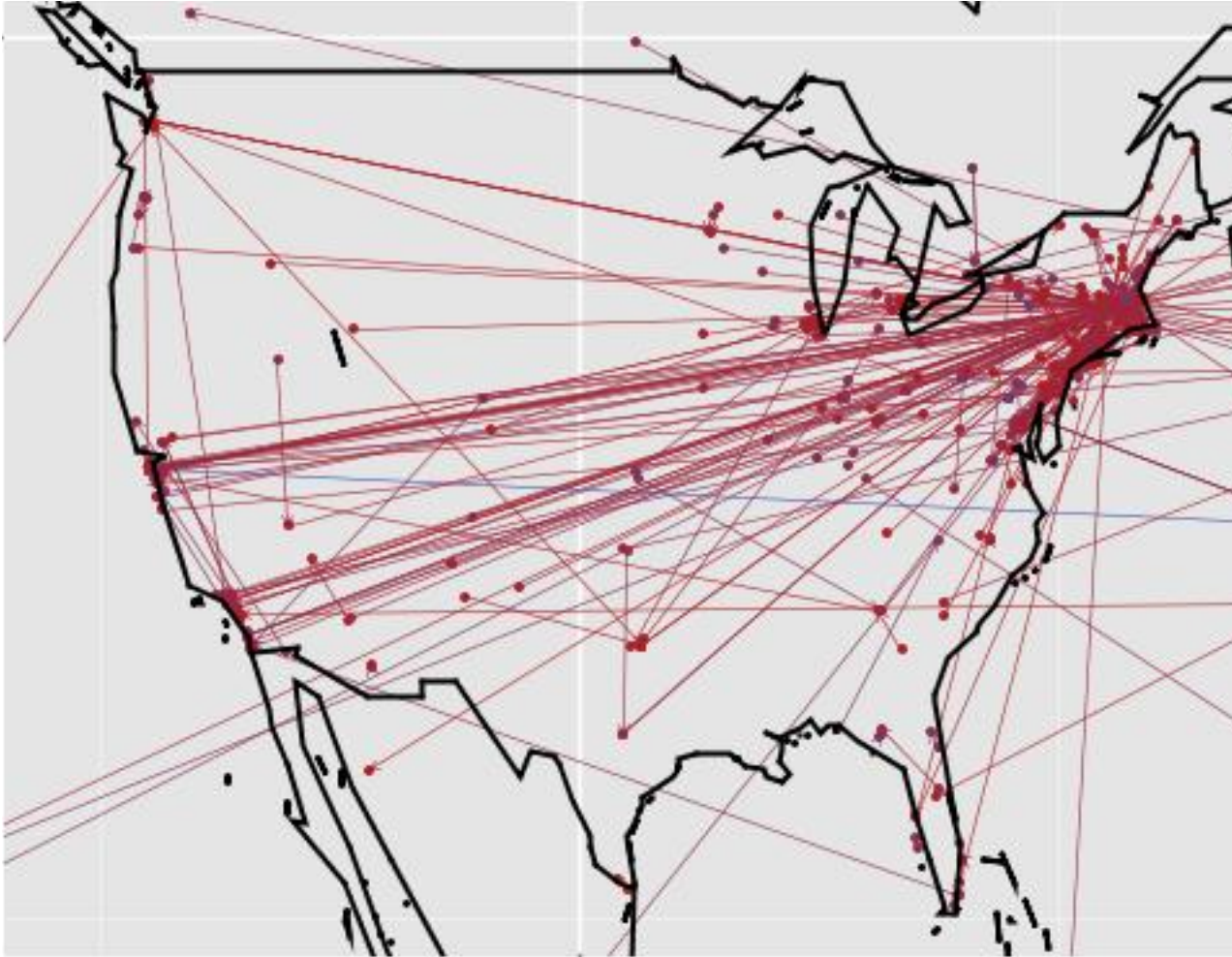
Sample Movie Factors

- ▶ Action-Thriller
- ▶ Horror-Supernatural
- ▶ Dystopia-Political
- ▶ Romance

Libraries : psych, mice



Inference – Distance Traveled



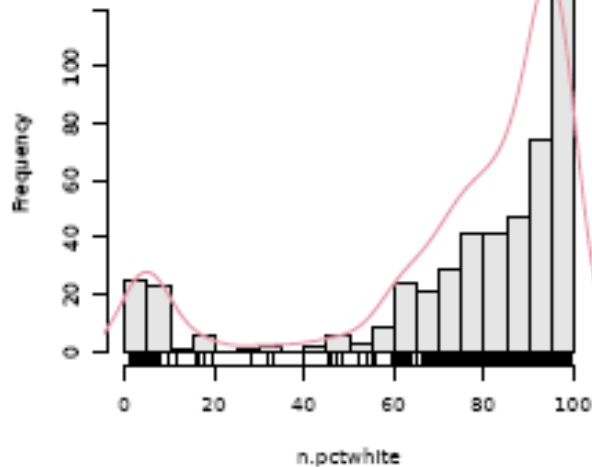
Libraries :
ggplot2,
geosphere,
maps



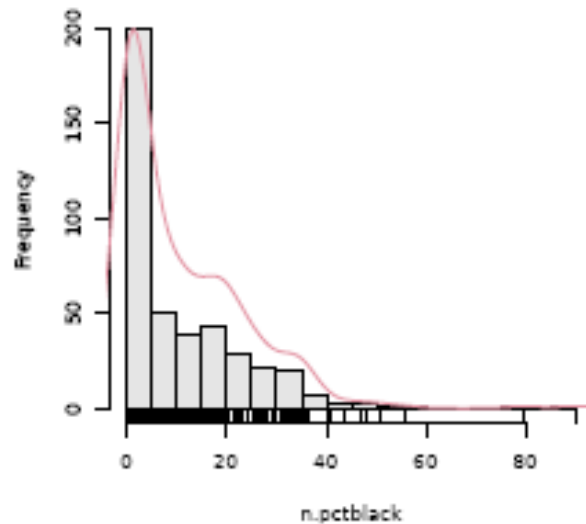
Inference - Race

name	rank	count	prop100k	cum_100k	white	black	api	aian	2prace	hispanic
SMITH	1	2376206	880.85	880.85	73.35	22.22	0.4	0.85	1.63	1.56
JOHNSON	2	1857160	688.44	1569.3	61.55	33.8	0.42	0.91	1.82	1.5
WILLIAMS	3	1534042	568.66	2137.96	48.52	46.72	0.37	0.78	2.01	1.6

Distribution of n.pctwhite



Distribution of n.pctblack



(Chang,2010)

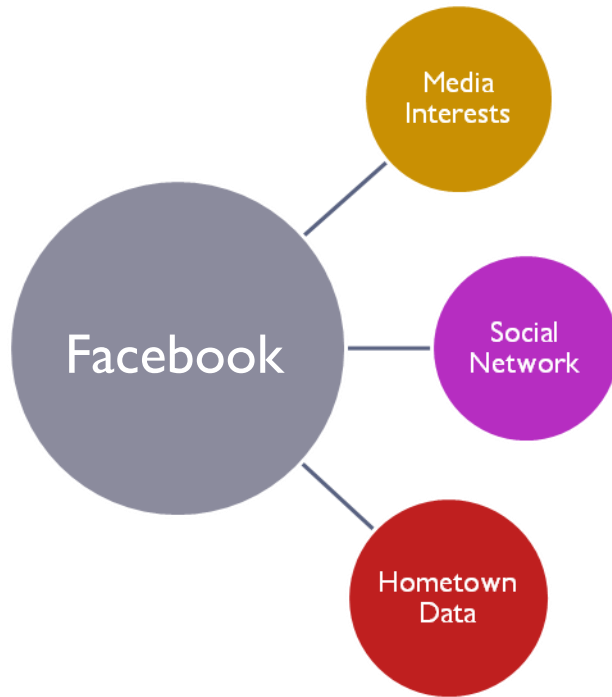


Inferences - Overview

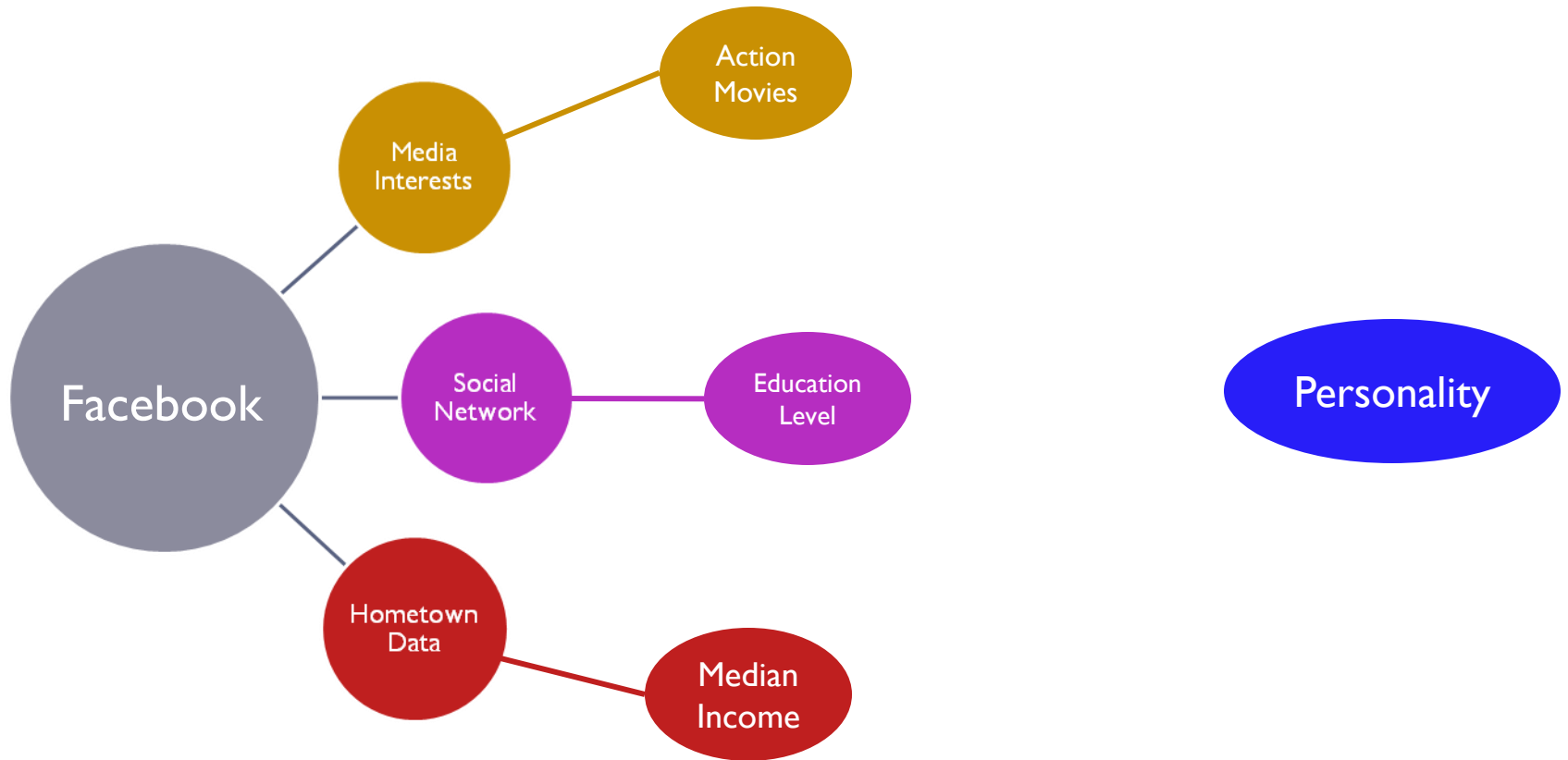
Type	Example variables
Base demo-graphics	Age, Sex, Education level, Geographic location, Ethnicity based on last name
Social Network	Friend count, Social network density,
Hometown Data	Mean income, Average education level
Media Interests	Movie genres such as comedy, romance, dystopia-political,
Behavioral Actions	Distance moved, Amount of Facebook interaction



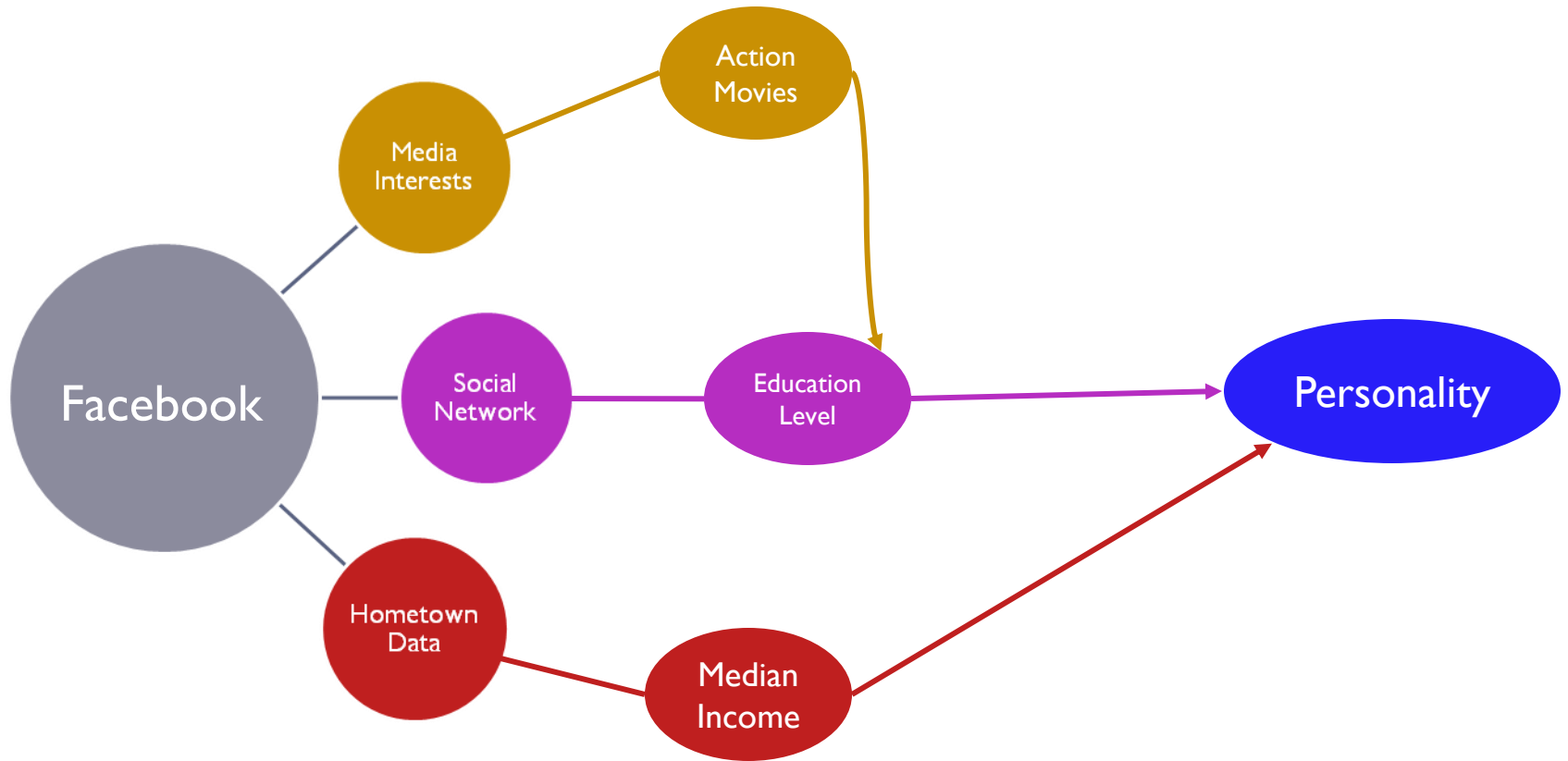
Inference



Inference

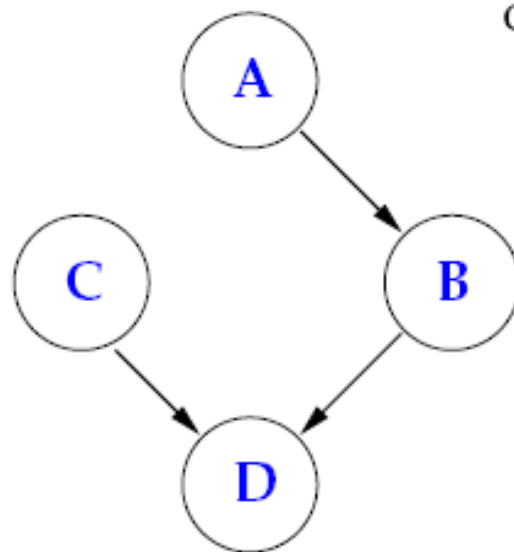


Inference



Bayesian Networks

- ▶ Directed graphical models representing joint probability distributions
- ▶ Edges represent conditional relationships
- ▶ A parent is related to its children and its children's children



CPT of B

	A='F'	A='T'
B='F'	0.23	0.6
B='T'	0.77	0.4

CPT of D

	B='F', C='F'	B='F', C='T'	B='T', C='F'	B='T', C='T'
D='F'	0.65	0.54	0.78	0.01
D='T'	0.35	0.46	0.22	0.99

(Wong, 2004)

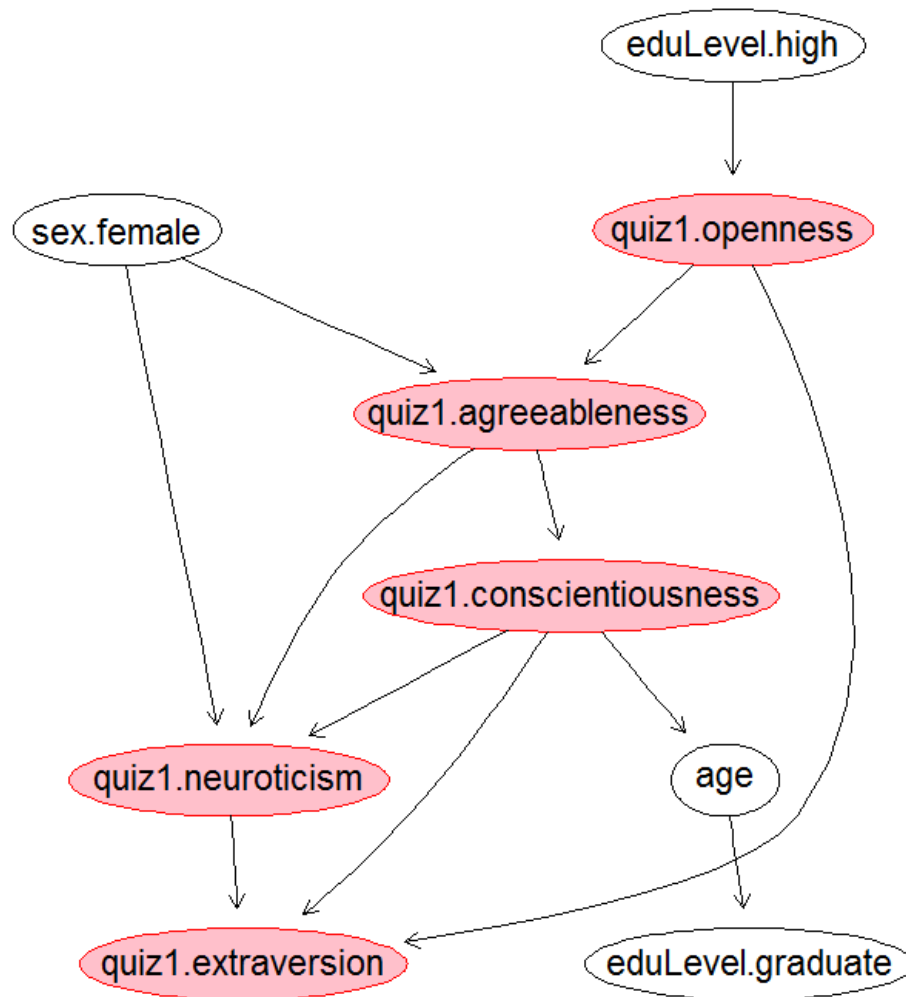


Bayesian networks

- ▶ Allow straightforward inference
- ▶ Belief propagation for limited evidence
- ▶ Clear underlying semantics
- ▶ Often have weaker predictive power than “black box” prediction methods



Bayesian Network - Example



Bayesian Network - Construction

- ▶ Primary goal maximizing predictive power of models on personality
- ▶ Continuous data + dummy variables
- ▶ Genetic algorithm used for variable selection
- ▶ Cross validation to prevent over-fitting
- ▶ Missing variables imputed with gibbs sampling
- ▶ Model's were assessed on the summed \mathbf{R}^2 of all 5 personality factors.
- ▶ Hybrid Bayesian Network construction used with the grow-shrink algorithm and hill climbing
- ▶ MLE for parametrization of Network

▶ Libraries : mice, bayestree, genalg, bnlearn, snow, foreach

Results – Current Model

Color By Source

Personality Data

Facebook Data

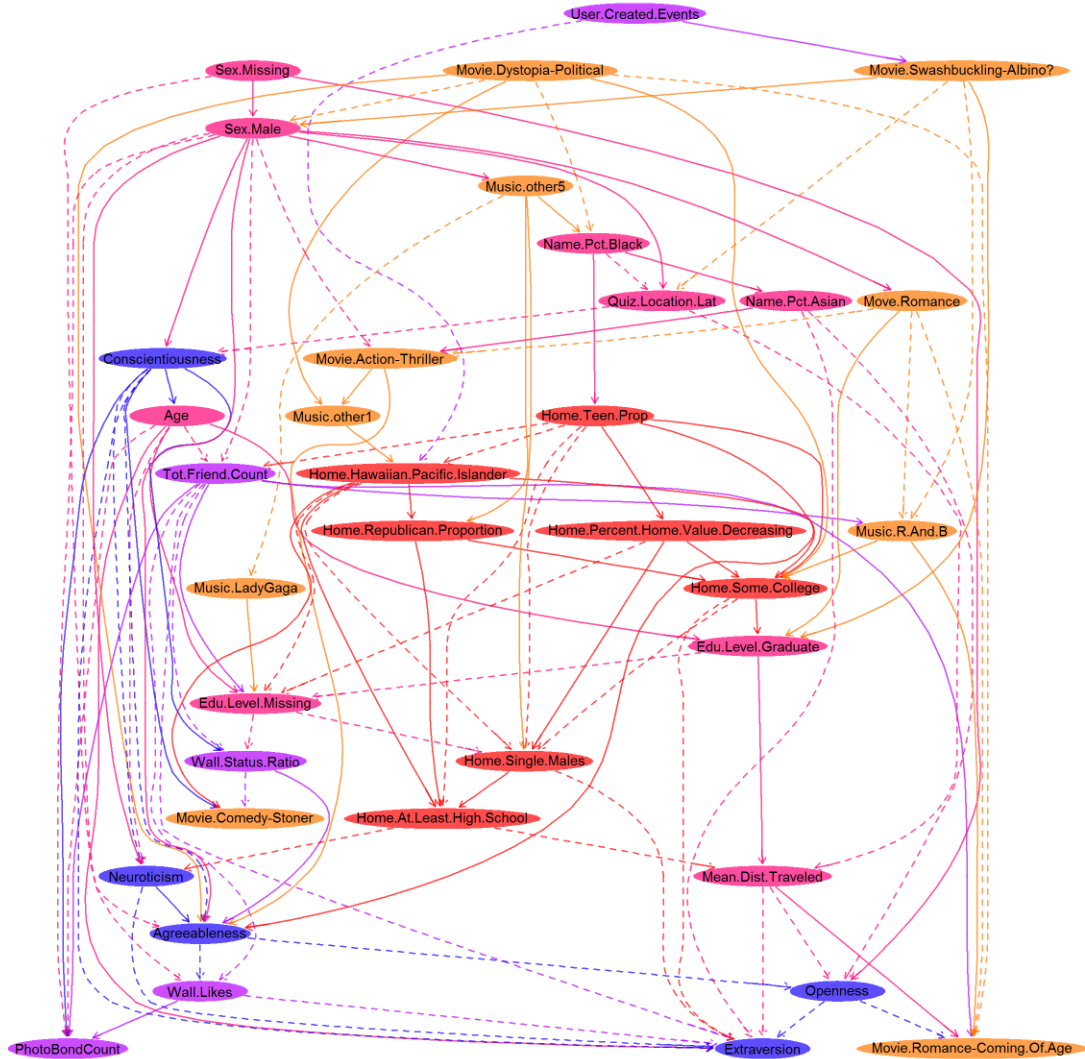
Hometown Data

Media Data

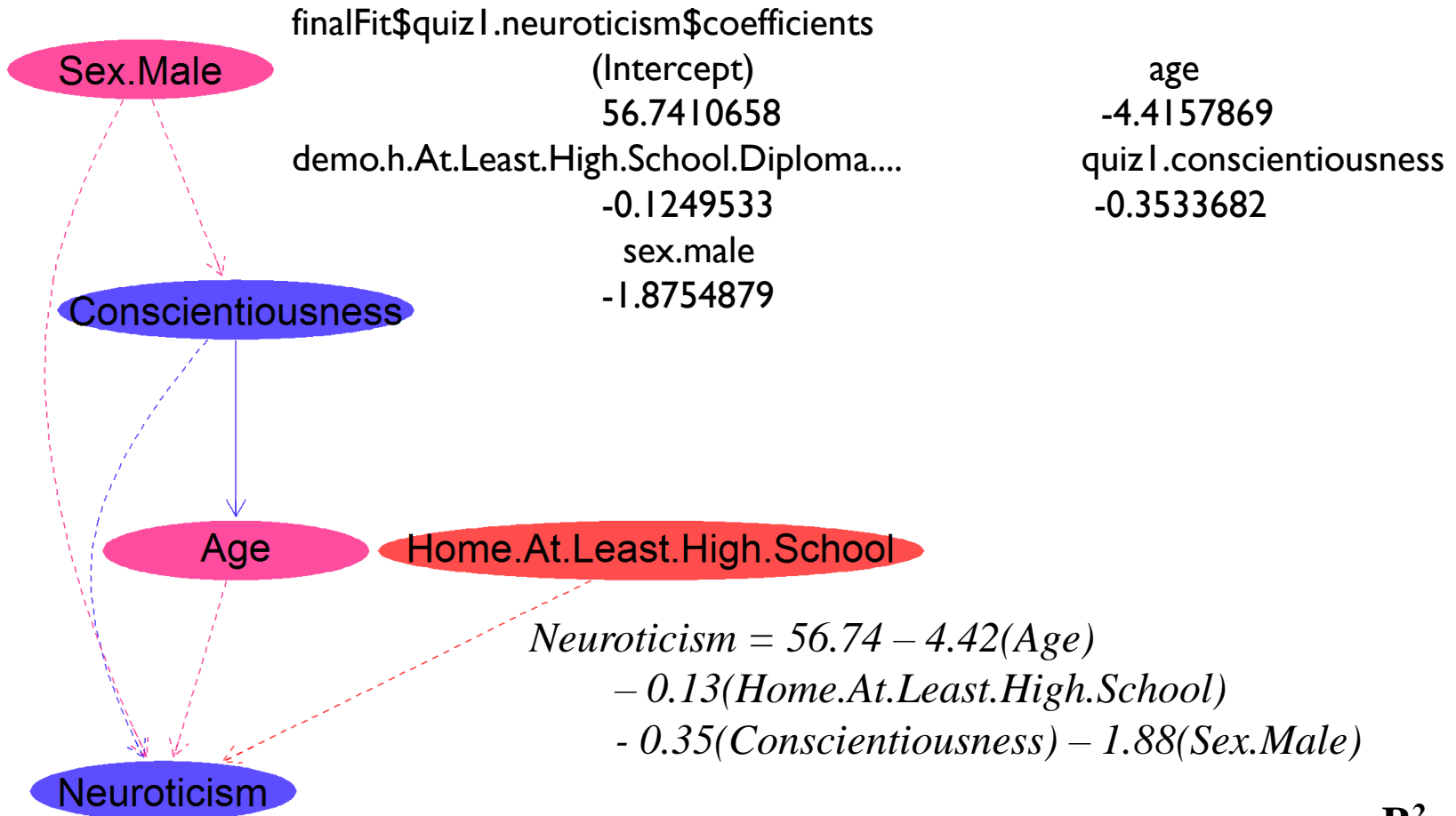
— Positive Relationship

- - - Negative Relationship

Libraries : RGraphviz



Results - Neuroticism



R² = .35

Results – Extraversion, Openness

$$\begin{aligned} \text{Extraversion} = & -2.81 - 0.59(\text{Mean.Distance.Traveled}) + 4.71(\text{Age}) \\ & + 5.86(\text{Conscientiousness}) + 2.49(\text{Wall.Likes}) \\ & + 32.04(\text{Home.Some.College}) + 3.76(\text{Home.Single.Males.city}) \\ & + 0.15(\text{Conscientiousness}) + 0.29(\text{Openness}) \\ & - 0.27(\text{Neuroticism}) + 1.53(\text{Name.Pct.Asian}) \end{aligned}$$

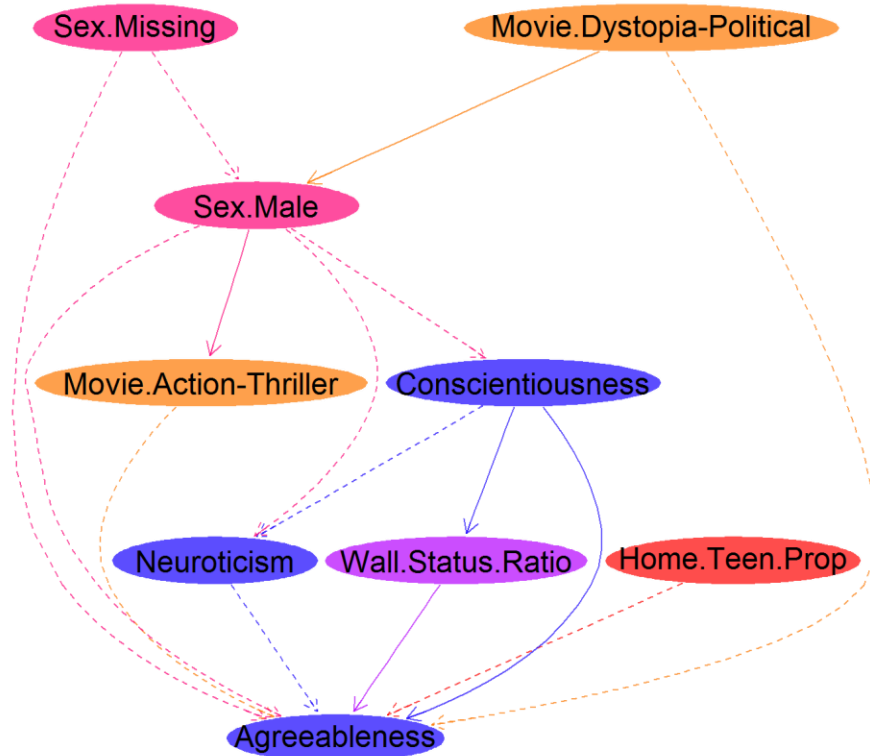
R² = .56

$$\begin{aligned} \text{Agreeableness} = & 42.38988 \\ & - 1.26(\text{Sex.Missing}) - 0.63(\text{Movie.Dystopia-Political}) \\ & - 25.99(\text{Home.Teen.Prop}) - 0.49(\text{Movie.Action-thriller}) \\ & + 6.51(\text{Wall.Status.Ratio}) + 0.08(\text{Conscientiousness}) \\ & - 0.29(\text{Neuroticism}) - 2.47(\text{Sex.Male}) \end{aligned}$$

R² = .46



Results - Belief Propagation



- ▶ Before evidence is introduced, model assumes marginal probability distribution
- ▶ Probability updated with evidence



Results – Belief Propogation

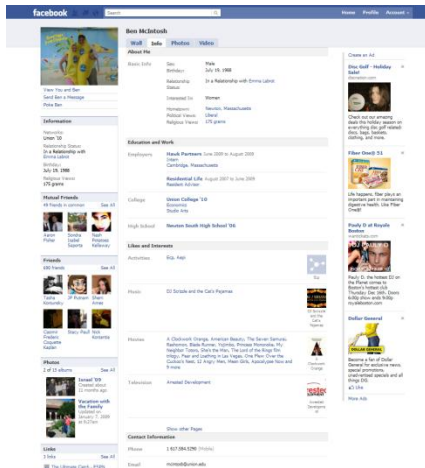


Factor Analysis

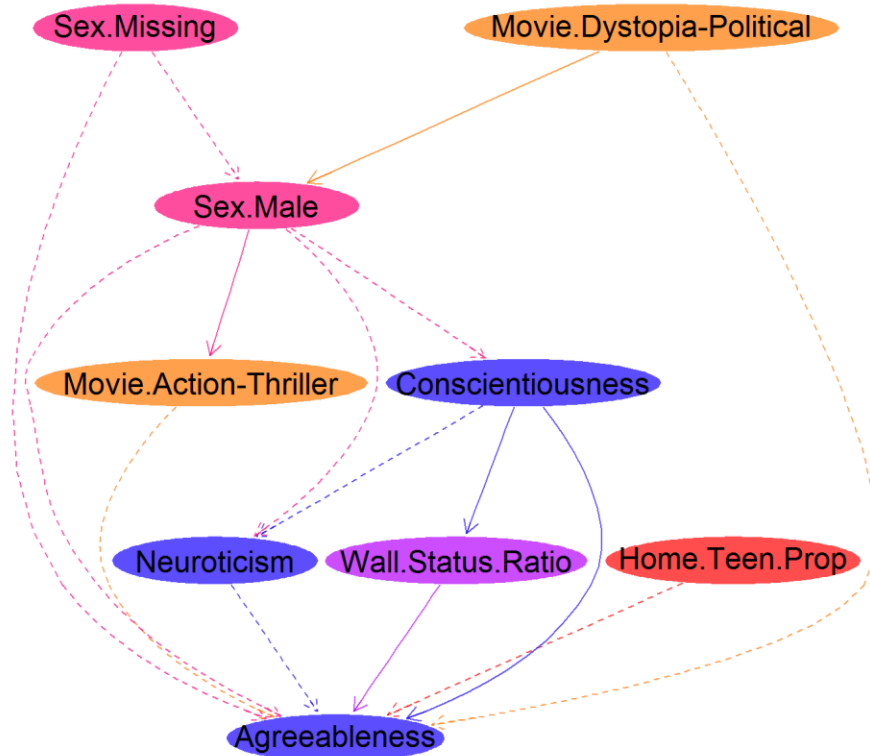


$Agreeableness = 42.38988$

- 1.26(*Sex.Missing*) - 0.63(*Movie.Dystopia-Political*)
- 25.99(*Home.Teen.Prop*) - 0.49(*Movie.Action-thriller*)
- + 6.51(*Wall.Status.Ratio*) + 0.08(*Conscientiousness*)
- 0.29(*Neuroticism*) - 2.47(*Sex.Male*)



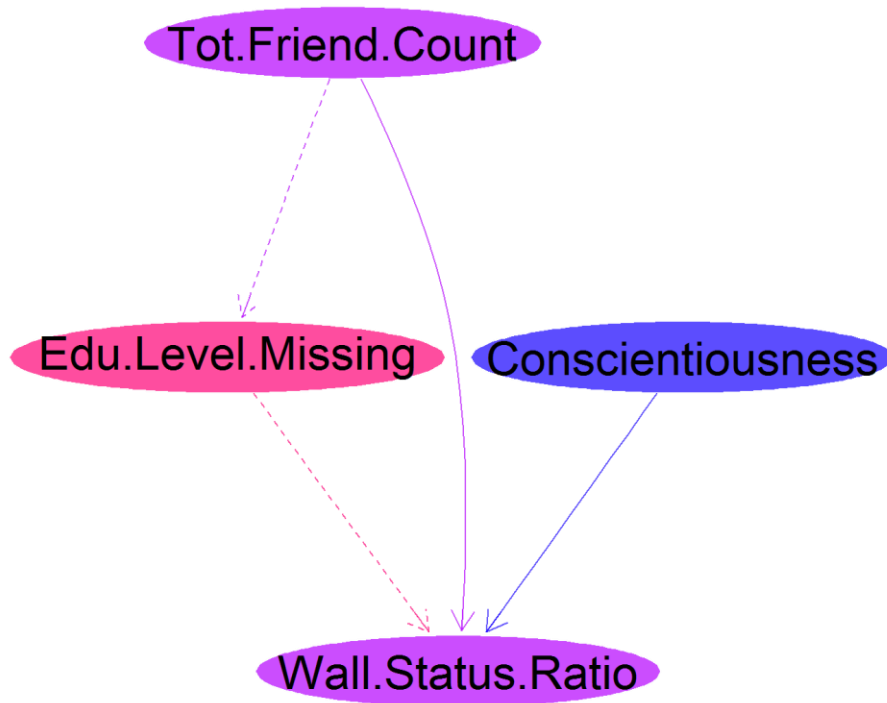
Results - Belief Propagation



- ▶ Belief propagation also generates predictions for non-personality variables



Results – Outward Prediction



- ▶ **Wall.Status.Ratio** is a ratio of users status updates vs. how much their friends post on their walls
- ▶ His attributes are used to predict actions, which then predict personality



Results – Weakness'

- ▶ Large sample size needed for discrete BN construction
- ▶ Sample collected from small area
- ▶ Relatively low Predictive accuracy



Further Findings

- ▶ Other modeling techniques can generate more accurate results
- ▶ Similarly, BN models which optimize a single personality measure are significantly more accurate
- ▶ SVM : Extraversion $R^2 = 0.84$
- ▶ Single Node Optimization BN: Extraversion $R^2 = 0.75$



Future Work

- ▶ Using personality for outwards prediction
- ▶ A larger sample size would allow more conditional inference
- ▶ Improving predictions by predicting single nodes at a time
- ▶ Using streaming data to update predictions with a Dynamic Bayesian Network
- ▶ Could be scaled to constantly update belief about individuals, their personalities and preferences.



References

- ▶ Ashton, M. C., Lee, K., & Goldberg, L. R. (2007). The IPIP–HEXACO scales: An alternative, public-domain measure of the personality constructs in the HEXACO model. *Personality and Individual Differences* , 42, 1515–1526.
- ▶ Chang, J., Rosenn, I., Backstrom, L., & Marlow, C. (2010). ePluribus : Ethnicity on Social Networks. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*. Washington, DC: AAAI Press.
- ▶ Korb, K., & Nicholson, A. (2003). *Bayesian Artificial Intelligence*. Chapman & Hall.
- ▶ Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks : An International Journal of Structural Analysis* , 330–342.
- ▶ Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R package. *Journal of Statistical Software* , 35 (3).
- ▶ Wong, M. L., & Leung, K. S. (2004). An Efficient Data Mining Method for Learning Bayesian Networks Using an Evolutionary Algorithm Based Hybrid Approach.



End

- ▶ Thanks to :
 - ▶ Revolution Analytics
 - ▶ Hampshire College school of Cognitive Science
- ▶ Questions?

