Visualization methods for genealogical and RNA-sequencing datasets

Ph.D. Thesis Proposal
Lindsay Rutter

Program of Study Committee:
Dianne Cook (Major Professor)
Amy Toth (Major Professor)
Heike Hofmann
Daniel Nettleton
James Reecy

May 16, 2016

Sommaire

1. Introduction
2. Visualizing genealogy
3. Boxes
4. Text
5. Tables
6. Images

My Background

- Education
  - B.S. in Bioengineering
    Pennsylvania State University (2003-2007)
  - Major in Bioinformatics and Computational Biology
    Iowa State University (2012-Present)
  - Minor in Statistics
    Iowa State University (2012-Present)
- Internships
  - Okinawa Institute of Science and Technology (Summer 2014)
  - MathWorks (Summer 2016)

Motivation

- Use visualization to explore data, check data quality, assess model diagnostics, and compare results across methods
- Problem : Limited choice of plots
    - Solution : Develop new plots
- Problem : Large datasets
    - Solution : Improve computational expense
    - Solution : Repair overplotting issues
    - Solution : Enhance pattern detection methods
    - Solution : Incorporate interactive graphics

Previous research

- **Static visualization software :** `ggplot2` (Wickham 2009), `GGally` (Schloerke et al. 2016), `nullabor` (Wickham et al. 2014), `ggbio` (Yin et al. 2012)
- **Interactive visualization software :** `GGobi` (Swayne et al. 2003), `tourr` (Wickham et al. 2011), `plotly` (Sievert et al. 2016)
- **General visualization :** Parallel coordinate plots (Inselberg 1985, Wegman 1990), Visual statistical inference (Chowdhury et al. 2015)

Previous research

- **Genealogical visualization :** `pedigree` (Coster 2013), `kinship2` (Therneau et al. 2015), `GraphViz` (Gansner and North 2000), `Cytoscape` (Shannon et al. 2003)
- **Gene expression visualization :** `explorase` (Lawrence et al. 2008), `limma` (Ritchie et al. 2015), `edgeR` (Robinson et al. 2010), `DESeq2` (Love at al. 2014), `RUVseq` (Risso et al. 2014)
- **Gene expression visual inference :** (Yin et al. 2013)
- **Biological clustering :** (Newell et al. 2013)

Problems

- Standard genealogical plots can be ambiguous
- Popular RNA-seq visualization tools are misleading
- Time and space constraints in large RNA-seq data

Thesis proposal overview

- (Chapter 2) Visualizing genealogical data
    - Goal : Create unambiguous genealogy visualization plots, adapt genealogical plots for large datasets, incorporate interactive genealogical plots
- (Chapter 3) Visualizing clustering analysis of RNA-seq data
    - Goal : Develop tools to visualize and interact with gene clusterings to determine genes of interest
- (Chapter 4) Visualizing significance tests of RNA-seq data
    - Goal : Develop tools to visualize, interact with, and permute differentially expressed genes from significance testing

Visualization methods for genealogical datasets

Genealogy

- Study of parent-child relationships
- Provides tools to better understand traits that arise in lineages
  - Desirable (disease resistance)
  - Undesirable (hemophilia)
- Can be represented visually

Current visual tools

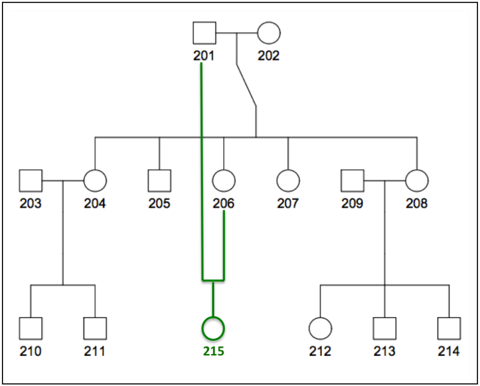

FIGURE: **kinship2** : Ambiguous position of green node, who is both second and third generation

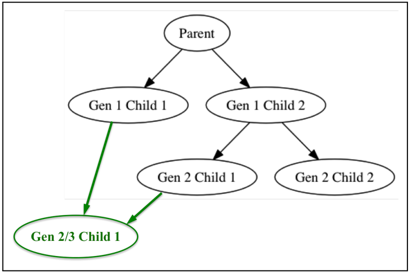

FIGURE: **Cytoscape** and **GraphViz** : Ambiguous position of green node, who is both second and third generation

ggenealogy

- **ggenealogy :**R package to visualize genealogical structures
- First example data is soybean genealogy
  - Soybean variety data collected from
    - Field trials
    - Genetic studies
    - USDA bulletins
  - Data frame of 412 rows (parent-child relations)
  - Each variety (n=230)
    - Developmental years
    - Copy number variants (CNV)
    - Single nucleotide polymorphisms (SNPs)
    - Protein content and yield

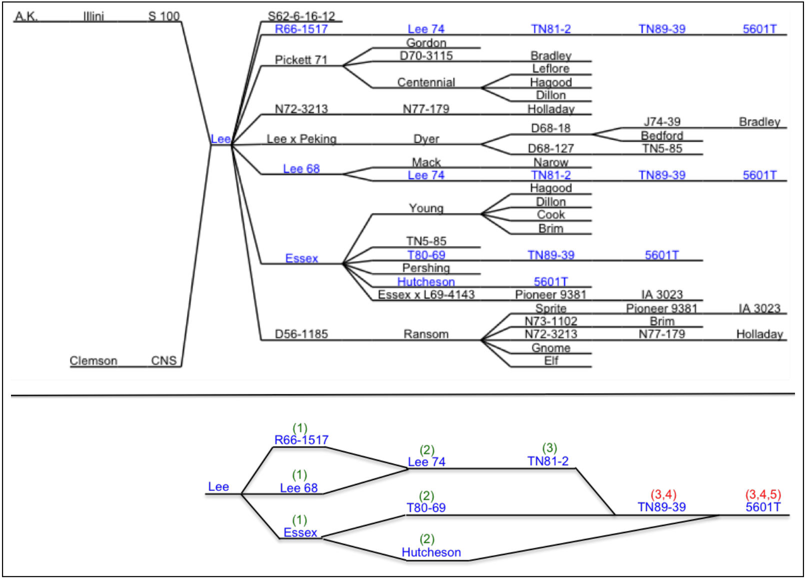

FIGURE: ggenealogy : Solution to ambiguous positions of nodes

Plot shortest path

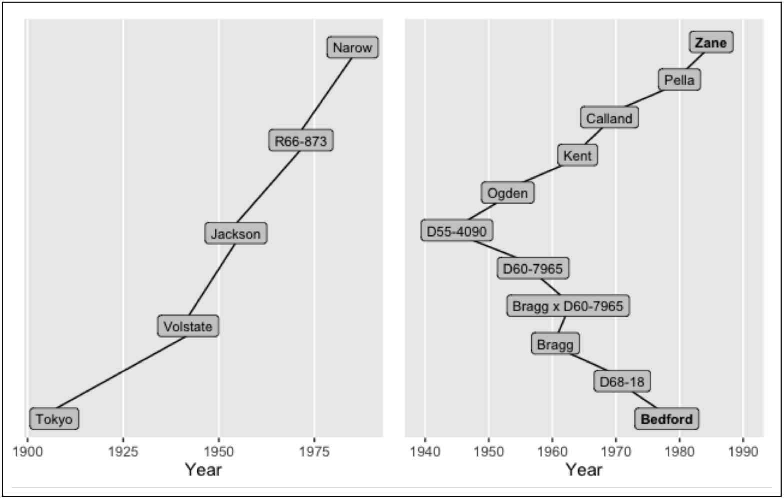

FIGURE: **Left :** The shortest path between Tokyo and Narow is composed of a sequence of parent-child relationships. **Right :** The shortest path between Zane and Bedford instead have a cousin-like relationship.

Superimpose path on full structure

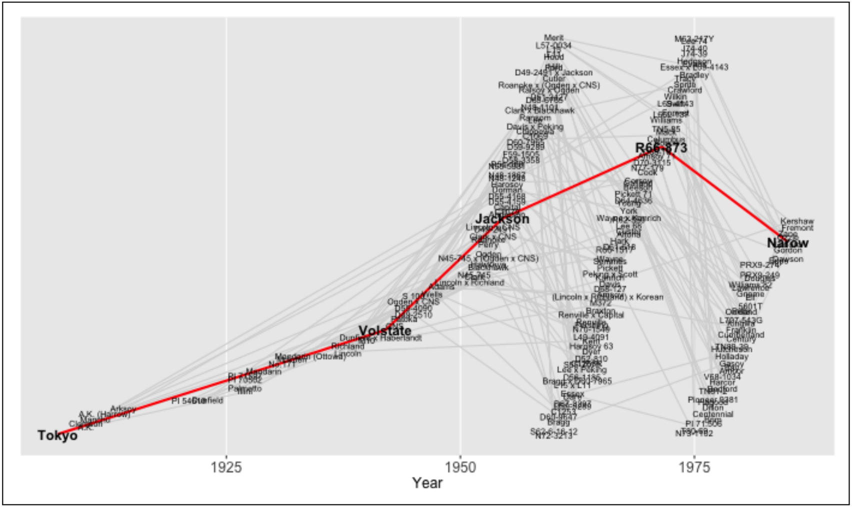

Figure: The shortest path between Tokyo and Narow, superimposed over the data structure, using a bin size of 3.

Superimpose path on full structure

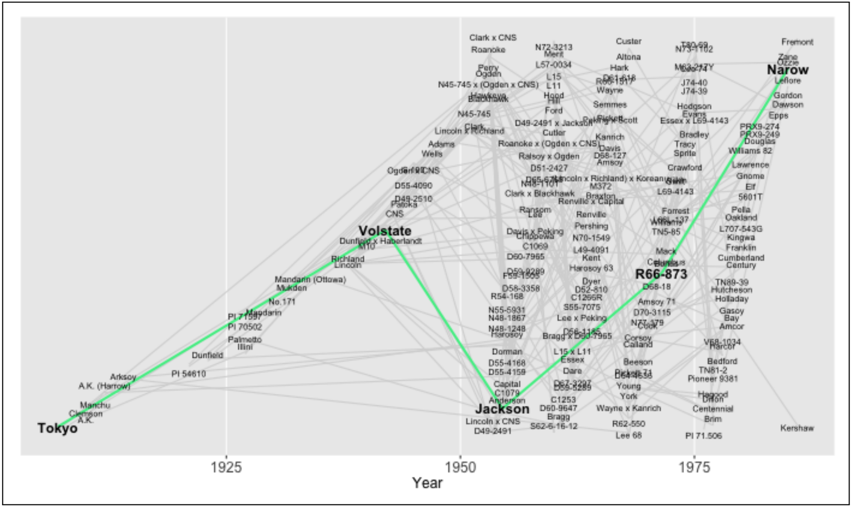

Figure: The shortest path between Tokyo and Narow, superimposed over the data structure, using a bin size of 6.

ggenealogy

- Second example data is genealogy of academic statisticians
  - Math Genealogy Project
    - Web database of genealogy of academic mathematicians
    - North Dakota State University Department of Mathematics and the American Mathematical Society
    - Queried for people with advanced degree in "Statistics" with parent with advanced degree in "Statistics"
  - Data frame of 8165 rows (3291 parent-child relations)
  - Each individual (n=7122)
    - Year of degree acquisition
    - Country of degree acquisition
    - School of degree acquisition
    - Thesis title

Including Code

```
> statISU <- statGeneal[which
(statGeneal$school=="Iowa State
University"),]$child
> length(statISU)
[1] 101
> numDISU <- sapply(statISU, dFunc)
> table(numDISU)
numDISU
 0  1  2  4 11 12 15 19
90  3  1  3  1  1 1  1
> which(numDISU == 19)
George Zyskind
```

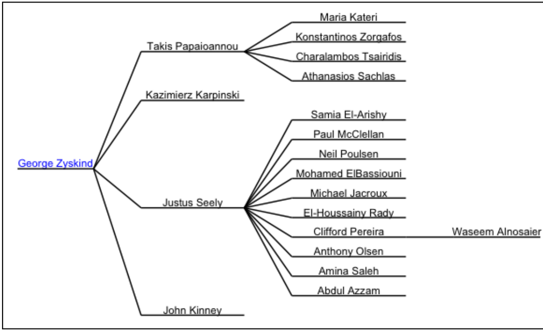

Figure: Nineteen "descendents" of Dr. Zyskind.

Including Code

```
> statUI <- statGeneal[which(statGeneal
$school=="University of Iowa"),]$child
> length(statUI)
[1] 54
> numDUI <- sapply(statUI, dFunc)
> table(numDUI)
numDUI
 0  1  7 25
48  4  1  1
> which(numDUI==25)
Edward Wegman
> which(numDUI==7)
Daniel Nettleton
```

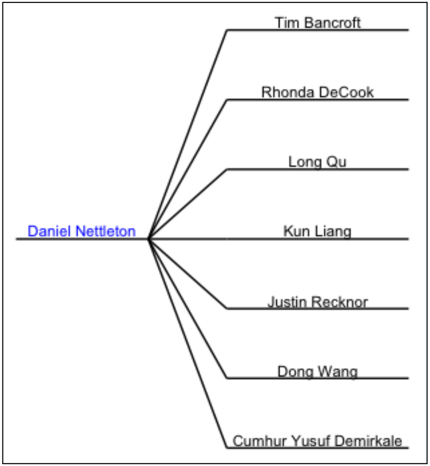

FIGURE: Seven "descendents" of Dr. Nettleton.

Blocks

Bloc simple

- Premier point
- Second point
- Troisième point

Bloc exemple

- Premier point
- Second point
- Troisième point

Bloc alert

- Premier point
- Second point
- Troisième point

Boxes

Ceci est
une boite jaune

Ceci est
une boite orange

Ceci est
une boite marron

Ceci est
une boite violette

Ceci est
une boite bleue

Ceci est
une boite grise

Titre de la frame

Voici du texte normal
Voici du texte `alert`
Voici du texte `exemple`
Voici du texte `emphase`

Tables

Couleur	Prix 1	Prix 2	Prix 3	Prix 4	Prix 5
Rouge	10.00	20.00	30.00	40.00	100.00
Vert	20.00	30.00	40.00	50.00	140.00
Bleu	30.00	40.00	50.00	60.00	180.00
Orange	60.00	90.00	120.00	150.00	420.00

Mon tableau des prix

Couleur	Prix 1	Prix 2	Prix 3	Prix 4	Prix 5
Rouge	10.00	20.00	30.00	40.00	100.00
Vert	20.00	30.00	40.00	50.00	140.00
Bleu	30.00	40.00	50.00	60.00	180.00
Orange	60.00	90.00	120.00	150.00	420.00

Tables

Couleur	Prix 1	Prix 2	Prix 3	Prix 4	Prix 5
Rouge	10.00	20.00	30.00	40.00	100.00
Vert	20.00	30.00	40.00	50.00	140.00
Bleu	30.00	40.00	50.00	60.00	180.00
Orange	60.00	90.00	120.00	150.00	420.00

Mon tableau des prix

Couleur	Prix 1	Prix 2	Prix 3	Prix 4	Prix 5
Rouge	10.00	20.00	30.00	40.00	100.00
Vert	20.00	30.00	40.00	50.00	140.00
Bleu	30.00	40.00	50.00	60.00	180.00
Orange	60.00	90.00	120.00	150.00	420.00

Titre de la frame

Figure: Éléments d'architecture bretonne typique du Sud de la France. (Wikipédia.fr CC-By-Sa)