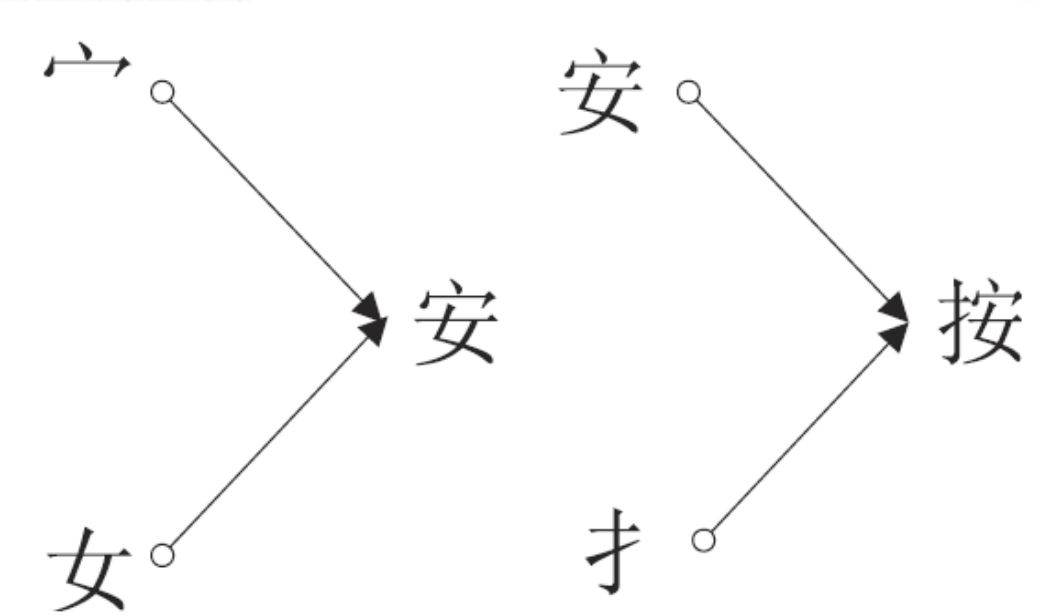
Using Complex Network Representation to Identify Important Structural Components of Chinese Characters for Foreign Language Students

Today's Presentation

- **Review of Chinese characters**
- Methodology

Chinese Characters

- One-syllable symbol
- wú wéi ér
  無 爲 而
  (“building blocks”)
- Radical(s) → Component(s) → Character



JianYu Li, Jie Zhou: Chinese Character Structure Analysis Based on Complex Networks (2007)

wikimedia.org

My Project

- Complex network of structural components
- Find hubs and clusters
- Analyze topology
- Suggest vocabulary learning system

Today's Presentation

- Review of Chinese characters
- **Methodology**

Composition Data (Network Structure)

- Wikimedia Commons

File format

1	2	3	4	5	6	7
龖	15	⿲	焱	6		思

1. Chinese characters / Unicode order.
2. Number of strokes in the character (not always reliable)
3. Composition kind (see below)
4. First character part (may be composed of several characters, if the composition does not exist as a single character).
5. ...and Number of strokes in this first char. part.
6. Empty = verification made ; "?" = still to do.
7. Second character part ; "*" when no different second part (primitives, or repetitions).
8. ...and nb of strokes for second part.
9. Verification of second part.
10. Cangjie codification (for easy sorting)
11. Radical (or * if the character itself is the key)

http://commons.wikimedia.org/wiki/Commons:Chinese_characters_decomposition

Character Frequency (Boundaries/ Weights)

- Frequency varies with:
  - Geography
  - Time period
  - Domain
- Most lists are for simplified characters

"Frequency and Stroke Counts of Chinese Characters"

- Chih-Hao Tsai, 1993-1994
- Usenet newsgroups
- 170 million+ characters in corpus
- 13,060 unique Big5 characters
- Large corpus, limited domain

"Hong Kong, Mainland China and Taiwan: Chinese Character Frequency – A Transregional, Diachronic Survey"

- Ho Hsiu-hwang and Kwan Tze Wan
- Nearly 4 million characters in corpus
- Variety of literary texts
- 3 different decades
- Smaller corpus, maybe more variety
- Exclude mainland characters (simplified)

"The Most Common Chinese Characters"

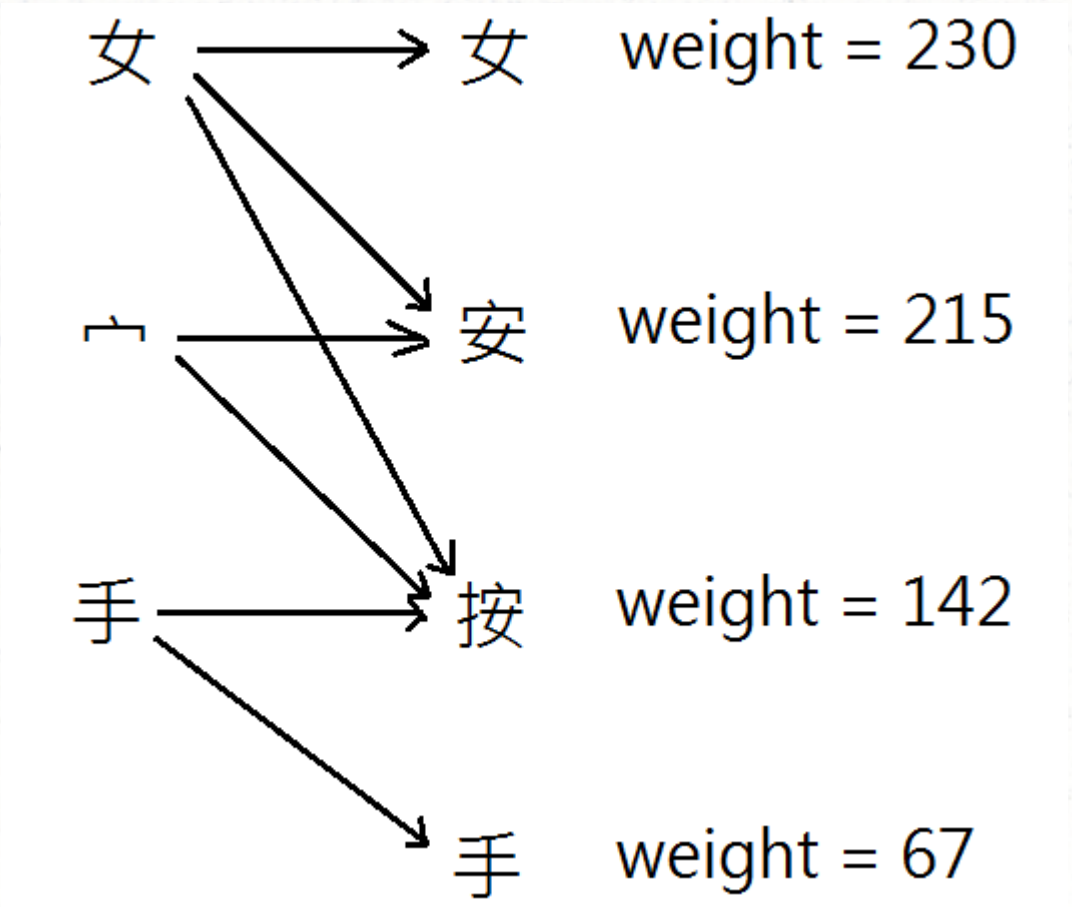
- Patrick Zein, online
- Based on dictionaries and other lists
  - Influenced by Jun Da's widely-used list
- For simplified characters
  - Traditional equivalents given
  - Equivalence is not perfect
- Only 3,000 unique characters

Data Preparation

- Composition table → NET format
  - Two network types:
    - Radical-based
    - Component-based
- Apply frequency list
  - Eliminate infrequent characters
  - Assign weights based on frequency

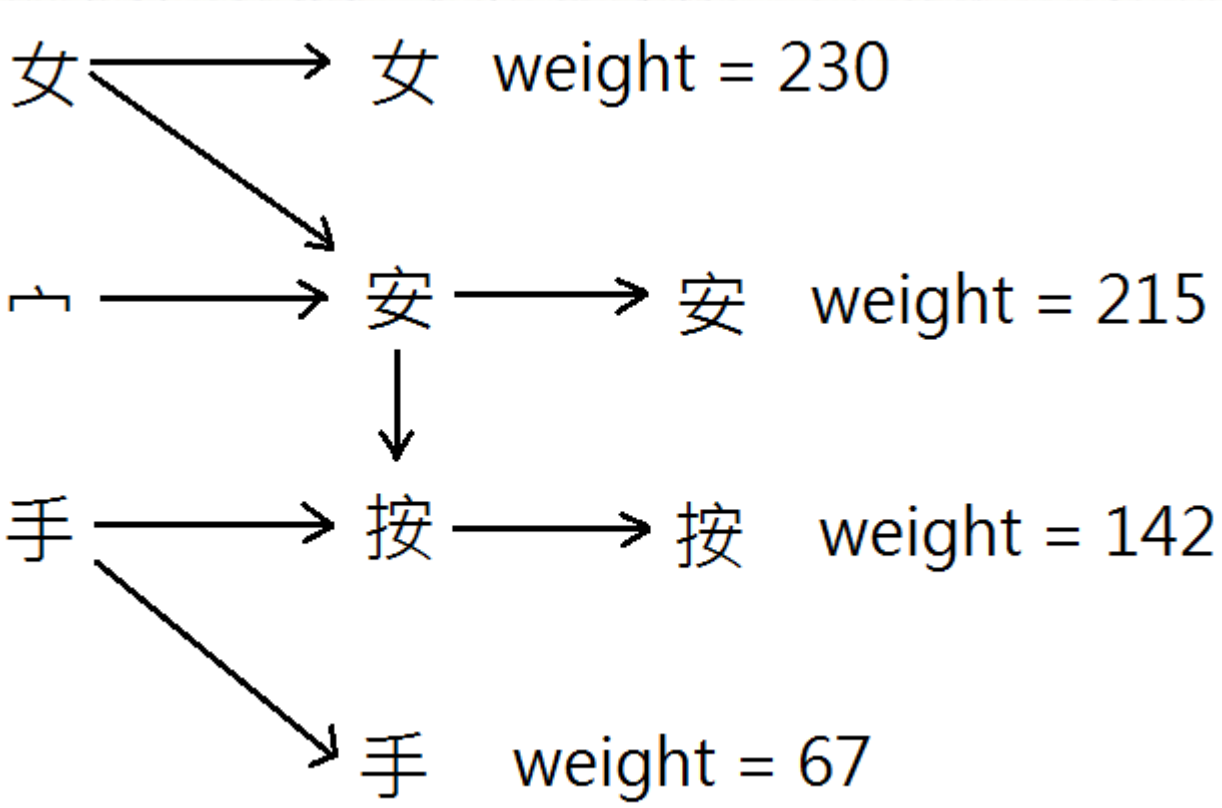
Radical-Based Network

Connects characters directly to all constituent radicals



Component-Based Network

Includes complex subcomponents



Network Metrics

- Diameter, Clustering Coefficient, etc.
- Weighted Degree Distribution
  - Identify hubs that are part of high-frequency characters
- Unweighted Degree Distribution
  - Identify hubs that are part of many different characters
- Radical-Based vs. Component-Based
  - Compare degree distributions

Cluster Identification

- Convert to unimodal format (remove character vertices)
- Find *k*-cores using different values of *k*
- Look for a value of *k* that yields many clusters of similar size
- Each cluster could correspond to a textbook lesson
- Reintroduce weights and rank clusters by total weighted degree

Network Variables

- Radical-Based vs. Component-Based
- Number of characters included
- Frequency list used

I will compare data from several different networks.

Goals

- Calculate basic network metrics
- Identify hubs and clusters
  - Find clusters that would work as textbook units
- Observe effects of variables
  - network size
  - frequency list
  - network structure