## C.4 Enron E-Mail Corpus and Network

Enron Corporation was an energy and commodity services company based in Houston, Texas. It began business in 1985 and grew to be one of the ten largest companies in the United States, achieving a stock price of $90.75 in mid 2000. By November 2001, however, its stock price fell to less than $1 a share. Enron filed for bankruptcy on December 2, 2001, prompting investigations by the U.S. Securities and Exchange Commission and the Federal Energy Regulatory Commission.

Enron's fraudulent practices became a matter of public record and brought corporate accounting and auditing practices into question. The Enron scandal provides a lesson in business ethics, well documented in popular books and articles.

Most of the executive employees of Enron and had nothing to do with the scandal, but their e-mail records became a matter of public record when the Federal Energy Regulatory Commission released them as part of its investigation. The Enron case data are Enron-centric—for non-Enron actors, we see only their communications with Enron executives.

As one of the few sources of real e-mail data in the public domain, the Enron e-mail archive and network represent a substantial opportunity for research in text analytics, online communications, and social networks. The current Enron e-mail corpus, occupying more than two gigabytes of storage and 500 thousand files, contains folders for 158 executives and over 200 thousand e-mail messages. The e-mail network consists of more than 36 thousand nodes and 183 thousand links.

McLean and Elkind (2003) and Eichenwald (2005) provided popular business books about the Enron scandal. Tim Grieve's *Salon* article quoted selected messages from the e-mail archive (2003), focusing on Ken Lay (1942–2006), CEO of Enron for most of the period from 1985 through 2002. Original source materials relating to the Enron case are available from the Federal Energy Regulatory Commission at `http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp`. The Enron e-mail corpus, maintained by William W. Cohen of Carnegie Mellon University, is available at `http://www.cs.cmu.edu/~enron/`. Data showing the from-node and to-node structure of the e-mail network data, drawn from Leskovec et al. (2009), are available as part of the Stanford Large Network Dataset Collection at `http://snap.stanford.edu/data/email-Enron.html`. An overview of the Enron data is provided by Klimmt and Yang (2004). The Enron data have been the source of many studies over the past decade (Leber 2013), with a special issue of *Computational and Mathematical Organization Theory* devoted to their analysis (Carley and Skillicorn 2005).