Using shared references to uncover social networks

Tracking the spread of ideas on the web is a key component of information gathering in many domains: the study of political and social movements, epidemiology, intelligence, marketing, and finance. Speaking broadly, there are two problems: (i) tracking an idea as it spreads from user to user and potentially changes ; and (ii) identifying the network of users and sites within which an idea persists. Knowing the answer to one of these questions can help us with the other. Knowing that a page belongs to a community of pages organized around some idea can make us more confident that we have correctly classified its content. Knowing that content of a page expresses an particular idea can provide evidence that we should we more heavily weight links to it in our community network. Problem (i) is generally addressed with the use of statistical models that classify content; problem (ii) with network-based methods which analyze link structure and traffic patterns. The goal of this research is to see whether we can combine these methods to produce (i) a more reliable content classifier; and (ii) a more reliable network analysis.

More concretely, can we build a system that simultaneously incorporates link, citation, and referential information in classifying pages, while it incorporates content information in identifying and weighting interpage references in mapping a community network?

In this short position paper, I will focus on the use of synergistic interactions between content-based and network-based approaches to identify political or ideological leanings. First, I will assume a rather broad definition of network link. Let A be the page being analyzed. A first-order link from page A to page B exists when (i) there is a hyperlink from page A to page B; (ii) page A mentions the author of page B; (iii) A mentions the organization responsible for page B. A second order link from A to B exists when A and B refer to the same person, text, event, or group. These are second-order links because we may imagine a virtual node in our network for each referent of a person, text, event, or organization; when A and B mention the same person, text, event, or organization, then there is a path of length 2 from A to B through the node.

Obviously recognizing second order links requires content recognition of a limited sort: we must identify bits of language identifying a particular person, text, event, or organization. Some of these content recognition problems are more challenging than others, but none is trivial. Even recognizing the referent of a proper name is not always easy, as the same entity may be evoked with a variety of linguistic devices. Ian Stuart Donaldson may be referred to in a single work as "Ian", "Ian Donaldson", "Ian Stuart Donaldson", "our beloved Ian", and "our founder". The form of a reference may be informative in itself, just as the text label on a hyperlink is. For instance, an outgroup newspiece on the Blood & Honour militant group would be much more likely to refer to Donaldson with a journalese title prefix, as in "Founder Ian Donaldson" than an ingroup document would.

But in extending the notion of link in this way we must take care not to create a graph too large and densely connected to be of any use. Again, a content-based idea is of use. Whether they are terrorist cells or anti-vaccine

crusaders, a key factor in the cohesiveness of doctrine-based communities is how well they are able to establish a sense of group identity, often in the face of a continuing outside threat.

Therefore one of the most important ways to track ideas is to identify the signals of group identity and the functions that maintain it. We hypothesize that a system of shared references is one important signal of group identity, and shared references may have both positive and negative poles. That is, a group system of shared references to persons, texts, events, or groups which have a strong positive or negative affect for the in-group. As an example take the case of militant groups in the U.S. The following table gives examples of positive and negative poles of each type:

	Positive	Negative
People	Ian Stuart Donaldson (ISD)	Barak Obama
Orgs	Aryans, Blood and Honour	mud people, Zog, New World Order
Texts	The Turner Diaries	Origin of the Species
Event	Founding of SS	Death of Hitler, Martyrdom of ISD
Pubs	National Alliance's Attack!	New York Times

Only shared references with high positive or negative group affect count in establishing group identity. How do we determine which those are? This is a content-based question, and there is a content-based approach. Each word and phrase in a document has a distribution vector telling us what other words it co-occurs with. The most salient other words are modifiers occurring at statistically significant rates. When strongly negative vocabulary co-occurs with some referring expression E at a significant rate, that is evidence that E has strong negative affect. The case for this negative affect being a strong signal of group identity is strengthened if we can find network-based evidence. That is, looking at other documents in our network of links, if the referring expression occurs at significant rate (versus outgroup documents), that is one kind of evidence. If its co-occurrence with negative vocabulary is significant (versus outgroup documents), that is another.

In sum, we have the elements of a bootstrapping process. A small number of ingroup documents may be used to provide links to find more. We now have a small network of sites. This community provides a body of texts sufficient to establish high-affect shared references. Shared references establish second-order links between sites, and a significant number of second-order links is as good as a first order link. Thus second-order links help us find more members of the ingroup, generating more documents, and so on. In this study, I illustrate this methodology using U.S. white militant groups as an example.