Review of "Trawling the web for emerging cyber-communities"
(by Vinay Goel)

The subject of the paper is the systematic enumeration of emerging / implicitly defined communities. The authors motivate a graph theoretic approach to locate such communities on the web and describe the algorithms and the associated algorithmic engineering.

The authors provide the reasons for systematically extracting these communities: they are good information sources, they provide insight into the intellectual evolution of the web and are useful for purposes of targeted advertising. An important reason that the authors fail to mention is that these communities can help researchers interested in a particular area of research discover/become aware of each other and hence avoid duplicating each others work.

These implicitly defined communities are characterized by the certain graph structures. They describe the process of scanning through a web crawl and enumerating these subgraph signatures on the web – 'trawling' the web.

The authors do not mention how they estimate the number of explicitly defined communities (mention citations / method used etc.) The paper makes a large number of forward-references to the method used and the results obtained. This should be avoided. Also, the relationship that their efforts have with the prior work in this field is not clearly described in every case. (e.g. the section on Metadata: do their efforts overcome problems? build on this method?)

Trawling exploits the structure of co-citation in the web graph to extract implicitly defined communities at their nascent stage. They provide the intuition that web communities are characterized by dense bipartite graphs and finding a community is then just a matter of finding (and then using) the core (complete bipartite subgraph) of a bipartite graph. In a core, the pages that contain the links as termed as fans, and the ones being referenced as centers. The syntactic definition used by the authors to describe a potential fan (has links to at least 6 different websites) is too limiting. Instead, the 'quality' of the links could / should have been taken into account.

The authors do a good job of highlighting the important problems of mirroring and duplication on the web and eliminate these problems by employing a shingling approach. To further reduce the size of the data being analyzed, they describe trimming the data by pruning the centers by in-degree (centers with large in-degree are pruned). Then the system works on this data by looking for cores (by iterative pruning and inclusion-exclusion pruning algorithms). These data algorithms have been designed in such a way that the running time grows linearly in the size of the output (highly desirable considering the scale of the web). The filtering of nepotistic cores is based on a very "loose" definition of "same web-site". The false negatives rate obtained following this filtering step of cores has not been mentioned. The process terminates with the extraction of cores of the desired characteristics.

The manual evaluation highlights the recoverability, quality and reliability of the results obtained; they present encouraging results. A mechanical method to evaluation would be much more feasible (as recognized by the authors too).

This paper makes a valuable contribution to the field of information extraction on the web (but the authors need to address some of the questions/issues raised in this review).