Water **2014**, *6*, 1601-1641; doi:10.3390/w6061601

OPEN ACCESS

water

ISSN 2073-4441
www.mdpi.com/journal/water

Article

Water Relationships in the U.S. Southwest: Characterizing Water Management Networks Using Natural Language Processing

John T. Murphy [1,*], Jonathan Ozik [1], Nicholson T. Collier [1], Mark Altaweel [2], Richard B. Lammers [3], Andrew Kliskey [4], Lilian Alessa [4], Drew Cason [5] and Paula Williams [5]

1 Computation Institute, University of Chicago, 5375 S. Ellis Avenue, Chicago, IL 60637, USA; E-Mails: jozik@uchicago.edu (J.O.); ntcollie@uchicago.edu (N.T.C.)

2 Institute of Archaeology, University College London, 31-34 Gordon Square, London WC1H 0PY, UK; E-Mail: m.altaweel@ucl.ac.uk

3 Institute for the Study of Earth, Oceans, and Space, University of New Hampshire, 8 College Road, Durham, NH 03824-3525, USA; E-Mail: Richard.Lammers@unh.edu

4 Center for Resilient Rural Communities, University of Idaho, 875 Perimeter Drive, Moscow, ID 83844, USA; E-Mails: akliskey@uidaho.edu (A.K.); lalessa@uidaho.edu (L.A.)

5 Resilience and Adaptive Management Group, University Alaska Anchorage, 3211 Providence Drive, Anchorage, AK 99508, USA; E-Mails: drewcason@gmail.com (D.C.); pwilliams@uaa.alaska.edu (P.W.)

* Author to whom correspondence should be addressed; E-Mail: johntmurphy@uchicago.edu; Tel.: +1-630-252-3453; Fax: +1-630-252-9559.

Received: 17 January 2014; in revised form: 14 May 2014 / Accepted: 15 May 2014 / Published: 3 June 2014

Abstract: Natural language processing (NLP) and named entity recognition (NER) techniques are applied to collections of newspaper articles from four cities in the U.S. Southwest. The results are used to generate a network of water management institutions that reflect public perceptions of water management and the structure of water management in these areas. This structure can be highly centralized or fragmented; in the latter case, multiple peer institutions exist that may cooperate or be in conflict. This is reflected in the public discourse of the water consumers in these areas and can, we contend, impact the potential responses of management agencies to challenges of water supply and quality and, in some cases, limit their effectiveness. Flagstaff, AZ, Tucson, AZ, Las Vegas, NV, and the Grand Valley, CO, are examined, including more than 110,000 articles from 2004–2012. Documents are scored by association with water topics, and phrases likely to be institutions are extracted via custom NLP and NER algorithms; those institutions

associated with water-related documents are used to form networks via document co-location. The Grand Valley is shown to have a markedly different structure, which we contend reflects the different historical trajectory of its development and its current state, which includes multiple institutions of roughly equal scope and size. These results demonstrate the utility of using NLP and NER methods to understanding the structure and variation of water management systems.

Keywords: water management; institutions; local *vs*. regional scale; local water suppliers; natural language processing; data mining; named entity recognition; network analysis

1. Introduction

The structural relationships among the public and private organizations that manage water can shape and constrain a population's responses to socio-environmental challenges. These institutions can vary in the spatial scale at which they are able to undertake action and can have competing or contradictory motives that restrict or even thwart cooperation or collective action. Management opportunities may exist at regional scales that are unavailable, unworkable or ineffective at local scales, while local-scale responses may be inadequate to address broader-scale problems and interests [1]. Mapping the structure of relationships among the institutions in a particular study area can, therefore, provide an important framework for understanding the range and scales of policies being adopted and actions being undertaken.

We present a method for the use of data mining on open and published information sources to recover the significant water management entities within a region and map their mutual relationships. We apply document classification strategies from natural language processing (NLP), information retrieval and named entity recognition (NER); we apply these to collections of published newspaper articles to identify institutional actors that have a role in the context of water management decisions in a given area. We then use document co-location to construct networks that map the relationships among those actors. Our study contributes to future efforts in which these structural relationships will be modeled and tested formally to understand their effects on water management actions.

Our approach is comparative: four different data sources from four different areas are considered. We structure our presentation by beginning with a theoretical background, then present a discussion of the four study areas, with an emphasis on how their historical trajectories have led to structural characteristics in the relationships among their water management institutions. We move to a discussion of our methods for data mining and network creation, including software that we have developed, and then present the results of these procedures. We close with an assessment of the results and a discussion of the potential applicability of the process in other study areas and future goals.

2. Background

Our theoretical framework has two central elements: the contention that the network of water management matters and the contention that public media sources contain a reflection of this structure,

which we can capture through a data mining technique and which we can then use for further, useful analysis. We present our theoretical background by discussing these two aspects in turn here and close with some comments about the advantages and limitations of our proposed technique.

2.1. A Structural View

Our orientation is structural. The prevailing system of water management in the U.S. Southwest is fragmented, in that it includes a myriad of local, state and federal agencies tasked with monitoring and making decisions regarding water access. For these institutions, authority in enforcing decisions often differs, and even where there are strong authorities involved, the jurisdiction of any agency is often based on geographic boundaries that do not correspond to watersheds [2]. The result is a condition in which multiple institutions must interact to manage diverse aspects of a single resource.

These interactions can be studied as an "ecology of games" [3,4] and, thus, as a situation in which each actor is attempting to achieve its goals in competition with, conflict against, or indifference to the goals of the other organizations with which it interacts. The "games" being played, however, may be less tractable than those found in traditional game theory, necessitating more flexible, but also more complicated forms of analyses, such as the "action situation", as proposed by McGinnis and Ostrom [5,6]. Others [7] have noted that the boundaries defining the water managers' interests are themselves subject to decisions by the water managers and that water managers actively create new network links. Governance by a suite of networked institutions can lead to challenging questions of authority and even legitimacy, as such networks can crosscut the boundaries of established democratic institutions [8,9].

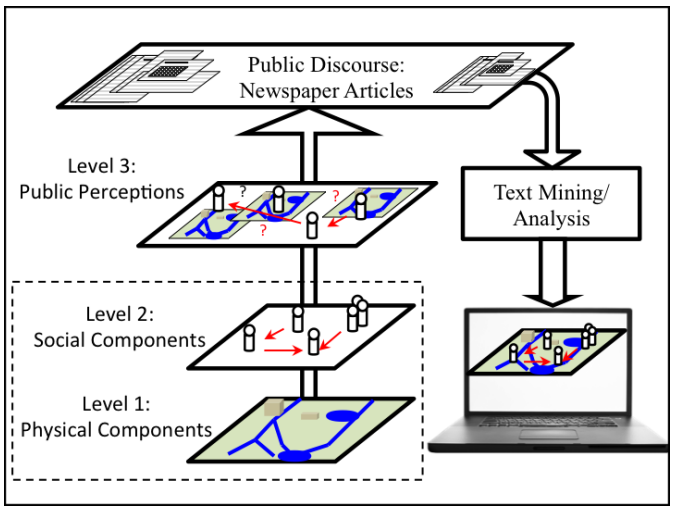
This situation can provide opportunities for inefficiency and for conflict. However, it can also carry benefits: from the view of adaptive management [10], the multiple roles and layers of interactions provide buffers against environmental threats and potential sources of innovation. Because of this, networks of governance institutions are recognized for their potential to meet challenges effectively, and they are therefore fostered and encouraged [11].

We are broadly interested in the question of which structures carry benefits or costs when faced with specific challenges or, alternatively, which structures facilitate adaptive change in the face of these challenges. There are three important characteristics that are salient in our approach. First, we do not presume that either a centralized or fragmented structure is inherently advantageous and, in fact, believe that each may be advantageous or vulnerable in their responses to different stressors. Second, our larger ambition is to model these systems, and ultimately to do so in conjunction with large-scale regional or global hydrological models. Our natural language and data mining project is in tandem with an additional project to create simulations of water management networks that can be linked to these existing hydrological models. Within this modeling project, we will be able to simulate an array of possible challenges and test the general performance of differently structured networks in the face of multiple challenges. Work of this kind can, we believe, provide a generalizable complement to real-world studies, which are necessarily limited to specific historical contexts and events. Finally, as part of our abstracted model, we do not characterize the nature of the relationships charted; more will be said on this below, but it should be made clear that network structures are emphasized over the relationships among nodes in this network.

2.2. Using Media as Public Discourse

Figure 1 illustrates our conceptual framework. There exist both a physical infrastructure and a social infrastructure to water management. These together comprise a network of relationships among water managers, and this network will have a particular structure; this structure may be fragmented, centralized and unified or be in some other arrangement. This network exists in an objective reality and can be observed and measured, though doing so may be costly and time-consuming.

Figure 1. Schematic diagram illustrating how physical infrastructure, public perception, discourse and analysis are used as a framework for identifying the water management structure.



There is, additionally, a public perception of this network; perceptions may or may not align with empirical realities of the network. This is a third "level" to our analyses, but it differs from both the physical and social infrastructure in that it is only a reflection of them. It is a heuristic that is employed to describe the collective beliefs and conceptions of many individuals.

Public perception can be thought of as a driver of the discourse that we will be mining. The analysis strategy adopted here attempts to extract from the discourse a representation of the empirical water management network in question. The key question is whether this representation matches the three levels given in Figure 1: whether it is commensurate with the perceptual/conceptual map held by the general public and whether it is commensurate with the physical and social infrastructure. The inaccuracies and lacunae of media reports in describing the empirical water management network can derive from multiple causes and can lead to systematic biases (e.g., [12]). Additionally, of course, the conceptual maps held by the public and the empirical networks may not match each other [13,14]. The feedback among the media, the public and the structure of networks of water governance has itself been the subject of study: management networks may crosscut the boundaries of existing democratic governance entities, and legitimacy must derive from the decision-making process (e.g., [9]); this process also may be impacted by the dynamics of the media through which the process, including the voices and roles of public stakeholders, is played out (e.g., [15]).

These issues are recognized in the analysis. However, we contend that it is possible to arrive at an accurate representation of the empirical structure despite working from texts that have been filtered

through an inaccurate conceptual map. There are advantages to be gained from working with a large number of documents: by examining repeated patterns across time and, when possible, across multiple sources, the data in aggregate may reveal a sharper picture than is found in any single document. While this is, of course, imperfect, it nevertheless can potentially alleviate a range of media errors.

2.3. Advantages and Limitations

There are several advantages to the approach worth noting here. First, the responses of populations to challenges related to water are impacted by their perceptions of these issues. Because of this, one important goal in studying water-related policy is assessing the perceptions of water issues held by the stakeholders and decision-makers (e.g., Gartin *et al*. [16]). Our study allows us to construe the physical and social realities within a given study area via perceptions of it held by the population under study, as these perceptions shape and are expressed in public discourse represented in public media.

Second, the approach is tailored to a convenient and useful scale. There exist some institutions that are involved in water distribution in a given area, but that are very small-scale providers (perhaps with only tens of customers; this is illustrated more completely in our methods section). We argue that these play little role in water management at the scales at which we are interested, and they generally also fall outside of our analysis given our data sources and techniques. Thus, our analysis strategy contains a built-in focus on entities that matter to our larger concerns. In this way, we may turn a potential liability, in the form of media bias, into an advantage. Note, however, that this advantage must be couched appropriately: for some kinds of analysis, attention to these small providers might be critical. It is a corollary to our modeling approach that we necessarily select elements to include in an analysis and those to exclude: no model includes all of reality, nor should it. In the case of this study, it is best to say that we expect an analysis at the scale we derive to be useful for a specific class of questions; many other questions, of course, could be raised that would require more or fewer components than our work generates. A study of the relationships between the larger and more central water management institutions and the smaller-scale institutions that our analysis excludes may be revealing, but we proceed with our analysis of the core institutions as the next step.

Finally, the approach can be performed rapidly, can be executed inexpensively and is generalizable across multiple contexts easily. Provided suitable data sets are available, a researcher can perform these analyses at low costs in terms of both money and time; moreover, the technique can be applied systematically to multiple areas, as is done in this demonstration and provide an efficient comparison of salient differences among multiple water management systems.

A few limitations of our approach should be borne in mind. First, the nature of the relationships among the nodes that the analytical process captures is not characterized. As will be shown below, a proxy for interaction (the co-occurrence of two entities in the same newspaper article) is used, but there is no attempt to differentiate interactions that result from conflict from those that result from cooperation. With certain difficulties and caveats, it is possible to begin to extract this information from the news sources: the documents in which two entities appear could be categorized based on their specific content, e.g., via sentiment analysis [17]. Such a characterization might be used to re-structure our networks, using, for example, a scale of cooperative *vs*. conflicting relationships to express the nature of the links between entities. It might in this way be possible to find that networks that are not

differentiated in the current study can be separated and revealed to have differing characteristics and performance in the face of specific challenges. Although we believe that the network in the abstract, without characterizing the relationships, can be analytically useful, we also recognize that the possibility of characterizing the network relationships offers great promise; however, for now, we leave this for future work.

An additional limitation of the work presented here is the restriction that the entire data set be considered chronologically "flat"; although the sources we will review cover as much as a decade, the analysis considers them all at once. This is best considered a limitation of the data utilized in the current study; the technique presented could easily be applied to subsets of the original data, divided into distinct periods and considered in sequence. However, dividing our data chronologically quickly introduced a problem of sparseness: small time slices yielded too few entities and links to be useful within our data set. This keeps us from seeing the evolution of networks through time, which would provide an intriguing additional dimension to our study. It might be possible to use a very coarse time scale (e.g., to compare the first half of the corpus with the second); more robust results, of course, could be gained by drawing from larger and/or longer data sets. It may also someday be possible to be more selective about links among institutions based on newspaper reporting, for example, by assessing the duration, impact and import of specific real-world events covered in sets of related articles rather than simply counting articles without regard to content. However, these possible avenues we also leave for future work.

Finally, there are known biases in the procedure toward specific kinds of institutions; there may additionally be unknown biases. Strictly speaking, the procedure recognizes two kinds of water management entities: those with names that match the extraction rules we have given and others that have explicitly and manually overridden these rules. In both cases, there may be a bias toward formal institutions over informal ones. A newspaper article in which a group of citizens debate some specific water management issue might not be recognized by our methods unless the group were referred to by some formal name; merely standing up in a town meeting, for example, would not be captured. In the extreme, a group that was reported on without being given a formal name in the newspaper articles would be extremely difficult to capture, but the less extreme case—a group that is referred to by a formal name, but that is either omitted by our automatic algorithm or manually omitted because the name does not convey the group's purpose—might also be omitted, potentially leaving out informal groups that might be of interest. Our hope is that the procedure we describe captures the majority of the actors at the scales of interest; however, the refinement of the technique to address these biases is also, clearly, possible.

It should be noted that the analysis employs a comparative approach and, further, that the work has been developed iteratively. Four areas are assessed that we believe, based on limited, high-level examinations, differ in their water management network structures. The same analysis method is applied to all four to generate maps of the water management networks derived for each area, and our claim is that the analysis method, applied in a uniform manner to the four data sets, generates different structures for each data set and that the structure generated via this method is, in each case, commensurate with what we know of the on-the-ground reality in the corresponding study area. This supports the argument that the method is genuinely capturing from the public discourse in media sources a reflection of the water management structure in each area. At several stages, the results have

been evaluated, adjustments made to the algorithm and the investigation re-executed. Excessive "tuning" was avoided, however, in favor of a general approach: adjustments made to the algorithm were applied to all four study areas and were made with the recognition that the changes would likely be generally applicable to all four study areas, as well as to novel study areas in future work.

3. Case Study Areas

The sample is drawn from four urban areas: Flagstaff, Arizona; Tucson, Arizona; Las Vegas, Nevada; and the Grand Valley, Colorado. These four areas were selected for two reasons: first, they all lie within the Colorado River Basin in the U.S. Southwest and, hence, present four windows into an area unified by a common set of problems, albeit each with its own position within that region. More pragmatically, each offered newspaper sources that were easily publically accessible; while data sources are increasingly common, many are also restricted or available at a high cost; for this study, we limited ourselves to sources that at the time we acquired them could be obtained without expense. The water management strategies pursued by the four areas have different historical trajectories and have arrived at distinct patterns of social and physical water management structures (Table 1).

Table 1. Study areas and characteristics.

Urban Area	Population (2010)	Water Sources	Water Management Structure
Flagstaff, AZ	65,870	Dams on Walnut Creek + increasing use of groundwater	Simple, centralized; isolated
Tucson, AZ	520,116 (Pima County: 980,263)	Aquifer + Central Arizona Project Canal (Colorado River)	Large, centralized; partial water importer
Las Vegas, NV	583,756	Lake Mead (90%), groundwater (10%);	Large, centralized; water importer
The Grand Valley, CO	146,723 (Mesa County)	Colorado and Gunnison Rivers (irrigation); groundwater from Grand Mesa and areas to E, SE (potable)	Complex; interlinked; communities

Water use in Flagstaff is almost entirely residential or commercial, not agricultural [18]. Flagstaff originally obtained all of its water from local surface sources, but beginning in the 1950s, this has been supplemented by groundwater. The fraction of groundwater has been increasing, and in recent years, more than 50% of Flagstaff's water has been drawn from groundwater sources, sometimes (*i.e.*, 1990 and 2002) approaching 100% [18]. Virtually all investments in water management infrastructure are undertaken by the City of Flagstaff, which can act essentially independently.

Tucson lies in a semi-arid basin along the Santa Cruz River. The river, a source for irrigation farming in historic, as well as in prehistoric times, no longer flows perennially, due to diversion of water for the upstream town of Nogales and other alterations. Tucson is situated over an aquifer from which residents today draw their water. Arizona's portion of the water from the Colorado, amounting to seven times that of Nevada's, is drawn from the Colorado at the town of Yuma and flows through canals and tunnels several hundred miles to Phoenix and Tucson. This system, called the Central Arizona Project (CAP), was established in the mid-1990s. Today, CAP water is mixed into the water in Tucson's aquifer.

The population in and around Tucson approaches one million; this is centered mainly within Tucson's city limits, which have expanded through time via a process of annexation. Although annexation has expanded the city's geographic and administrative extent and infrastructure, the provision of water is centrally controlled [19]. Independent "isolated" systems exist, but are under the aegis of Tucson Water [20]. Today, the greater Tucson area includes a small number of adjacent communities; but only one of these (Sahuarita) has a long history, and it lies at some distance from Tucson and has its own water source; the rest are communities a few miles outside of Tucson that have appeared as the population of Tucson itself has grown in the last two decades. The result of this historical process is that Tucson, despite being much larger than Flagstaff, is similarly dominated by a single water provider.

Las Vegas is marked by its transformation from a small railroad town in the early 1900s to one of the fastest-growing urban areas in the U.S. The Colorado River Compact, negotiated in 1922, apportioned Colorado River water among the states through which the river flowed. Because of its small population, Nevada received rights only to 300,000 acre-feet per year (0.37 km^3 yr^{-1}). Currently, Nevada has the right to draw more water, because of "credit" earned by implementing reclamation methods and through agreements with other Colorado River states, but the total amount it draws from the Colorado is still lower than the amount consumed by Las Vegas. The shortfall is supplied from groundwater wells that tap aquifers below the city [21].

Motivated by intensifying water shortages coupled with the expectation of continued growth, beginning in 1991, Las Vegas restructured its water management systems. The Southern Nevada Water Authority (SNWA) was created "for the purpose of acquiring and managing water resources for Southern Nevada, constructing and managing regional water facilities, and promoting responsible water use" [21] and subsumed seven independent water companies that had operated in the greater Las Vegas area, assuming many of their functions. In a short time, its functions expanded even further, eventually taking over water negotiations with other states previously performed by the State of Nevada [22–26]. The SNWA has been a key player in a recent longer-term strategy of trying to acquire the right to tap aquifers below the Great Basin areas north of the city [23].

The fourth study area, the Grand Valley, Colorado, sits at the confluence of the Colorado and the Gunnison Rivers and is home to several distinct communities; the best known today is the City of Grand Junction, which shares the valley with Palisade, Redlands, Fruita, unincorporated Clifton and a few smaller settlements. This has led to a complicated water management landscape. The early occupation of the valley by European settlers beginning in 1880 saw the construction of irrigation canals that quickly grew in size, length and capacity [27]. According to Simonds [27], as early as 1902, a primary diversion canal drawing water off the Colorado River and along the north edge of the valley was discussed; this became the Grand Valley Project, (GVP), the first large-scale work by the Bureau of Reclamation. Upon completion, pre-existing public and private water-related organizations were brought to the table and joined by new organizations that formed in response to the new opportunity the GVP offered. These were given rights to operate different sections of the resultant irrigation structure, and these distinct institutions, with different constituencies and different rules of operation, remain in place today.

Drinking water for the residents of the Grand Valley comes from other sources and is managed separately. Originally, each town in the valley sought its own supply. In the early 1900s, the City of Grand Junction began to purchase rights to water on the Grand Mesa to the south and east of the valley [28]. The other communities, including some additional water jurisdictions added in previously

unincorporated areas, have been forced to deal with more limited supplies and to seek water rights from more distant locations.

The outcome of this history is that control of water in the Grand Valley is divided among several different groups: at least four water districts and six irrigation districts exist, even though as the towns have grown to fill the valley, their systems have become more interconnected. These groups can cooperate [28]; in some cases, however, they choose not to and, instead, retain independent rights, as well as the accompanying costs and risks of remaining independent [29–31].

3.1. Summary and Expectations

Three of the study areas have arrived at comparatively centralized water management arrangements. Flagstaff's small population and isolation enabled it to be managed by only a few providers, even in its early history, and now, a single utility manages its groundwater sources. Tucson has grown much larger than Flagstaff, and the municipal provider has steadily acquired most of the other water providers, leaving a single utility in a dominant position, though still with a few comparatively small competitors within the city and some smaller, but primarily peer institutions in neighboring towns. Las Vegas, with multiple providers just over two decades ago, actively subsumed them under a single auspice. The Grand Valley is the exception. Having grown from several peer towns that now form a single metropolitan area, it still includes a number of independent institutions, all of comparable size; the division between irrigation and potable providers and the number of different sources available contributes to the fragmentation of the water management landscape.

We contend that this will be represented in the public discourse of the areas under study. In centralized areas, most of the discussion of water will involve the central entity. Discussions that involve an alternative entity (e.g., a neighboring town of Tucson) may take place and may be related to the central entity, but discussions that involve more than one of these peer entities are less likely. Conversely, in a fragmented area, such as the Grand Valley, discussions involving multiple entities will be more common. These may reflect disagreements, negotiations, conflict, cooperation or other interactions. We propose that this is a reflection of a situation in which there are overlapping domains and competing motives or interests and, thus, may represent a challenging context for water management. The network of these discussions as extracted by our analysis should include multiple network nodes that are all equivalently central, as shown by appropriate statistical analysis.

4. Methods

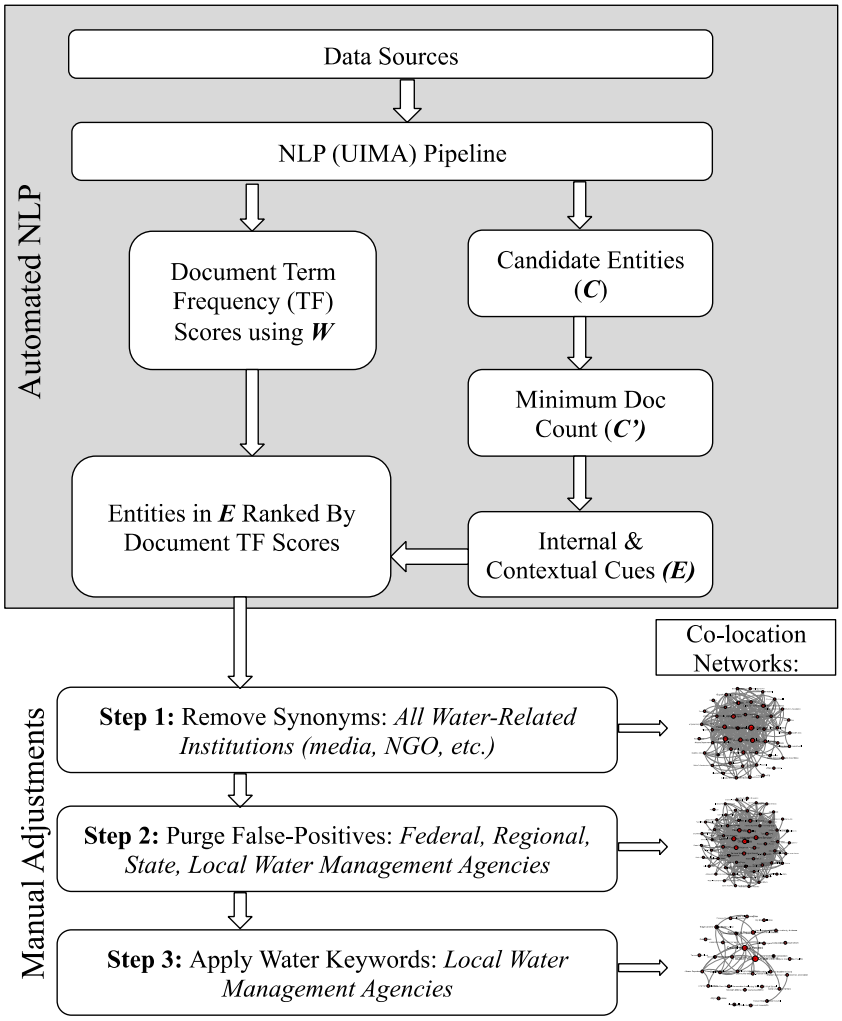
Figure 2 gives a very high-level overview of our analysis process; more detail is given below. Additionally, we provide the complete code for our software tool at our project website [32] and in the supplemental materials archived with this document.

4.1. Acquiring Data Sources

The automated data gathering process was broken down into two components. The first involved using the querying capabilities of the four regional news sources web sites to obtain links to articles. Because our interest is in local issues, articles that could be identified as nationally produced (e.g.,

syndicated) were ignored. The links were stored in index files along with the relevant meta-data for each link (e.g., news source, date, title). The second component involved following the links in the collected index files and extracting the content of each news story into local storage for analysis and processing by the natural language processing (NLP) "pipeline" as seen in Figure 2. This pipeline was constructed using the Unstructured Information Management Architecture (UIMA) framework [33]. As the documents move through this pipeline, a sequence of "annotators" adds markup to them; annotators can add markup indicating information about the properties of individual words (or "tokens") or phrases within the document (or about the entire document). Figure 2 shows the source data moving into the NLP pipeline and two results coming out of this analysis.

Figure 2. Analysis workflow diagram. NLP, natural language processing; UIMA, Unstructured Information Management Architecture.



4.2. Establishing Document Relevance

The objective of the left branch exiting the NLP pipeline in Figure 2 is to identify documents that are related to the general topics of water and water management. To do so, we calculate a statistic that measures the degree to which each document is related to these topics and then rank the list of documents in decreasing order on this value.

The NLP pipeline yields two outcomes that are used for this. The first is establishing the length, in tokens, of the document . The second is the identification and accumulation of words in the document that are found in a dictionary of water-related terms; this lexicon comprises the following list of terms, designated here using a boldface **W** per mathematical set notation:

Water Keywords (***W***): aqueduct; aqueducts; bodies of water; body of water; brook; brooks; canal; canalization; canalize; canalization; canalize; channelization; channelization; creek; creeks; flood tide; floods; H2O; irrigate; irrigates; irrigating; irrigation; lake; lakes; levee; levees; marsh; marshes; marshland; ocean; oceans; pond; ponds; puddle; rain; rainfall; river; rivers; sea; seas; swamp; swampland; swamplands; swamps; swamps; water; water channel; water channels; water conservation; water pollution; water regulation; water regulations; water supplies; water supply; water system; water systems; waters.

Each document in our collection has a title, which is extracted from our data sources. Repeated examples suggested that titles can be highly indicative of an article's topic; therefore, if the title includes one or more water-related terms, the raw count of water terms in an article is multiplied by 1.5 times the number of water-related terms in the title. The resulting value, called the term frequency (TF), is then scaled for document length (because longer documents may be expected to have more water-related terms merely by chance); this is achieved by dividing the original TF by the square root of the number of tokens in the document. While different specific formulas exist for TF and its scaling [34], the general approach is widely applied for retrieving relevant texts in a searched corpus [35].

4.3. Extracting "Water Authorities"

The right branch in Figure 2 traces the algorithm for identifying water authorities. This algorithm is a variant of named entity recognition (NER), which includes several steps involving different extraction strategies; these work together to arrive at a list of phrases that are likely to be the names of "water authorities". As is common in NER, special lexicons are used that inform the extraction process; the special lexicons that are used at this point in our NER procedures are:

Days of the Week (***D***): Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday;

Months (***M***): January, February, March, April, May, June, July, August, September, October, November, December;

Personal Titles (***T***): Administrator; Attorney; Chairman; Chairwoman; Chaplain; Chief; Columnist; Commissioner; Congressman; Congresswoman; Councilman; Councilwoman; Councilmember; DA; Deputy; Director; Dr.; Editor; Engineer; Examiner; Executive; Father; General; Gov.; Governor; Judge; Justice; Mayor; Manager; Marshal; Mgr.; Mr.; Mrs.; Ms.; Miss; Officer; Pfc; Pfc.; President; Prince; Prof.; Professor; Representative; Rep.; Reverend; Rev.; Secretary; Sen.; Senator; Sgt.; Sgt; Sheriff; Speaker; Superintendent; Supervisor; Tem;

Non-NNP Prepositions (***N'***): above; along; at; around; because; between; by; during; from; in; near; off; outside; since; through; toward; until; with.

The first step of the analysis is undertaken as a separate task within the NLP pipeline described above. This task is a custom annotator that creates a list of phrases that are potential water authorities, or "candidate water authorities", which we denote as set ***C*** (see Figure 2). This relies on annotators that mark the part-of-speech (POS) for each token and demarcate sentence boundaries. Our algorithm to

create set ***C*** proceeds through the sequence of tokens within each sentence; it is assumed that water authority designations will not cross sentence boundaries. As each token is encountered, its POS is adjusted to account for idiosyncrasies of the POS tagger (part of the OpenNLP package; see [36]) and to further the specific goals of finding candidate water authorities. The adjustment includes overriding the POS of any token that has been identified inappropriately (with respect to our purposes) as "noun or noun phrase" ("NNP") to another custom POS. The adjustments also include overriding the NNP designation from certain tokens that are presumed not to be of interest, such as the names of days of the week (***D***) and months of the year (***M***). Personal titles (***T***) are marked with a special POS designation. Additionally, some prepositions that rarely occur in water authority titles (***N'***) are distinguished from other prepositions.

As it proceeds through the document, the algorithm initially disregards all tokens that are not marked with the adjusted POS as NNP. Once an NNP token is encountered, it begins to assemble a sequence of tokens; the sequence continues so long as the token encountered has an adjusted POS of either NNP, "DT" (definite article) or "IN" (the normal preposition designation). Only sequences that have at least two tokens are recorded; the list of these sequences becomes the set of candidates, ***C***.

Our procedure (now outside of and using the results, including aggregate results, from the NLP pipeline) next eliminates any phrase in ***C*** that appears in fewer than two valid documents in the corpus, where validity consists of having at least one water keyword (***W***); this eliminates entities that are in no way related to water and further eliminates a large number of parsing errors, which are usually idiosyncratic and appear only once. The product of this filtering is a reduction of ***C*** to a new set that we term ***C'***. The algorithm next examines the entries, *c*, in ***C'***, considers each *c* as a potential entity and attempts to determine if it is likely to represent an institution. Following a common NER strategy [37], the algorithm examines *c*'s internal characteristics, as well as contextual cues from the words and phrases around each instance of that entity in the corpus. These two criteria provide a collection of features that are used to evaluate whether the phrase should be included in the list of candidate institutions or not. Several more lexicons are employed at this stage:

Administrative (***A***): Administration; Agency; Alliance; Authority; Board; Bureau; Department; Dept; Dept.; District; Center; Chamber; Club; Commission; Committee; Conservancy; Cooperative; Division; Foundation; Institute; Institution; Inst.; Office; Organization; Org.; Section; Service; Society; Soc.; Works;

Initial Geographic Feature Terms (G_i): Arroyo; Lake; Mount; Rio;

Geographic Features (G_f): Arroyo; Basin; Bay; Beach; Canal; Canyon; Cliffs; Coast; Creek; Cut; Dam; Desert; Divide; Falls; Foothills; Forest; Fork; Gap; Gorge; Gulch; Harbor; Lake; Lakes; Meadow; Meadows; Mesa; Mount; Mountain; Mountains; Ocean; Overlook; Peak; Peaks; Peninsula; Plateau; Pond; Pueblo; Range; Reservoir; Ridge; Rim; Rio; River; Rock; Park; Pass; Sea; Slope; Spring; Springs; Trail; Tunnel; Valley; Wash;

Geographic Terms (G_t): North; Northeast; East; Southeast; South; Southwest; West; Northwest; Northern; Southern; Eastern; Western; Northeastern; Northwestern; Southeastern; Southwestern; Upper; Lower;

Legislative (***L***): Act; Bill; Compact; Code;

Street (***S***): Avenue; Ave; Ave.; Boulevard; Blvd; Blvd.; Ct; Drive; Highway; Parkway; Pkwy; Rd; Rd.; Road; St; St.; Street;

Representative (***R***): Official; Officials; Representative; Representatives; Spokesman; Spokeswoman; Spokesperson;

Organization (***O***): Administration, Authority, Bureau, Co., Company, Department, Dept., Division, Service, Society.

The set of features we considered is given in Table 2, along with precise definitions for each. Note that for contextual features, even a single occurrence in any document in the corpus that matches the specified criterion will give the associated phrase a "true" value for that criterion.

Table 2. Internal and contextual criteria used for named entity recognition.

Feature	Symbol	Definition	Value
Count of terms	l	Number of tokens in the phrase	Integer
Count of non-NNP terms	n	Count of terms in phrase that appear in *N'*	Integer
Count of initial geographic terms	g_i	Number of tokens in G_i that are found contiguously at the start of the phrase	Integer
Count of end geographic features	g_f	Number of tokens in G_f that are found contiguously at the end of the phrase	Integer
Includes water	W	Phrase includes one or more tokens from *W*	T/F
Includes admin	A	Phrase includes one or more tokens from *A*	T/F
Ends with street	S	Final token in phrase is a member of *S*	T/F
Ends with legislative term	L	Final token in phrase is a member of *L*	T/F
Preceded by definite article	D_b	Contextual: token in document immediately preceding phrase is a definite article	T/F
Preceded by personal title	T_b	Contextual: token in document immediately preceding phrase is a personal title (from *T*)	T/F
Has representative	R	Contextual: will be true in two cases: (1) the word or words immediately preceding the phrase is "for", "from", "of", "for the", "from the" or "of the", and the word immediately before this is from *R* or *T*; or (2) if the word immediately following the phrase is from *R* or *T*, with the exception of "Dr." (which, in this, position usually means "Drive")	T/F
Followed by acronym	U_a	Contextual: true if the token immediately following is two or more capital letters enclosed in parentheses	T/F

Using the symbols defined in Table 2, the algorithm applies a set of criteria to move from ***C'*** to a reduced set of entities, ***E***, that are likely to be institutions. This set, ***E***, includes: all phrases more than three words long ($l > 3$); all phrases with non-NNP tokens in them ($n > 0$); all phrases that include administrative terms (A); all phrases that have representatives associated with them (R); all phrases that are preceded by definite articles (D_b); and all phrases that are followed by acronyms (U_a). This set is then reduced by removing all phrases that are preceded by personal titles (T_b), all phrases that end with street terms (S) and all phrases that end in legislation terms (L). The set is then further reduced by removing all those phrases that do not have a representative associated with them, but that either: (1) end in final geography terms (g_f) and include non-NNP terms; or (2) consist of a series of initial geography terms followed by one additional term ($l - g_i = 1$). This last reduction specifically addresses the difficulty of distinguishing between two kinds of "place" terms: those that are simply geographic

features and those that are names of places, such as towns or cities, and that, therefore, may also be discussed as if they are actors (e.g., "Mount Pleasant builds new treatment facility"). In set-builder notation, our entity recognition algorithm is:

$$\begin{aligned} \boldsymbol{E} = (c|((\mathrm{l} > 3) \vee (\mathrm{n} > 0) \vee \mathrm{A} \vee \mathrm{R} \vee \mathrm{D_b} \vee \mathrm{U_a}) \wedge (\neg \mathrm{T_b} \wedge \neg \mathrm{S} \wedge \neg \mathrm{L}) \\ \wedge \neg (\neg \mathrm{R} \wedge (((\mathrm{g_f} > 0) \wedge (\mathrm{n} > 0)) \vee ((\mathrm{l} - \mathrm{g_i}) = 1))) \end{aligned} \quad (1)$$

4.4. Candidate Institutions and Document Relevance

To arrive at a list of water-related institutions, we combine the document relevance calculated in Section 4.2 with set ***E***; in Figure 2 this is found when the right and left branches from the NLP pipeline merge. The approach taken to accomplish this merger is to take the average TF score for the documents in which a given candidate appears. More specifically, for each candidate, we rank the documents in which they appear by TF score and take the average of the top N of these. N is limited because some institutions that are very highly water related are also active in other domains and so appear in a large number of non-water-related documents. For all the analyses presented here, N = 10.

Once each institution has a score that reflects how closely it is associated with water-related documents, the institutions can be ranked: those scoring more highly are more likely to be water management institutions than those at the bottom of the list.

4.5. Targeted Manual Revisions

Having completed the fully automated part of our procedure (indicated by the gray box in Figure 2), the algorithm moves to a phase where targeted manual revisions are used to eliminate errors that the automated procedure cannot detect. The two most significant kinds of errors that can affect the process described here are the inability to recognize synonyms and the existence of false positives. The problem of synonymy arises, because there can be multiple ways to refer to any given real-world entity. Our algorithm as constructed does not include an automated way to recognize that two different sequences of tokens, e.g., "City of Grand Junction" and "City of GJ", actually refer to the same thing. The solution is to allow the user to specify that any appearance of one term be treated as if it were the other term. The record in our database of the appearance of the synonym is modified such that all analyses proceed as if the canonical term had been observed in the data source instead of the synonym; all of the internal aspects of the canonical term are used in place of those of the synonym, but all of the contextual cues from each instance of the synonym remain unaltered.

The problem of false positives is resolved by allowing the user to specify that a particular entity is to be ignored. This allows the analyses to proceed as if no instances of that entity had been recorded.

One other datum is allowed to be altered manually by the user: the specification that a given candidate water authority includes a water keyword from ***W*** that can be manually set to either true or false. While this could be applied for theoretical reasons (some water terms useful at one point in the analysis may be counterproductive at another), the *de facto* purpose is to allow specific entities to be intentionally included or excluded from the final step in the analyses described below.

These manual corrections are deployed in a progression of stages, corresponding to a useful strategy that an analyst might use in a wide range of study areas. After first performing the basic entity

extraction and the ranking of candidate institutions by association with relevant documents, we move forward with the analysis in three steps. In practice, these steps are guidelines; the actual process can be iterative, allowing the analyst to re-start from Step 1 if appropriate, given further information revealed in a later step.

4.5.1. Step 1: Remove Synonyms

We initially pass through the data set and identify certain entity names that are virtually certain to be synonyms. An automated routine aids this by finding all pairs of entities that match, except for a final term, e.g., "Tucson Water" and "Tucson Water Department"; the terms used are those from set ***O***, as given above. The automated component is merely an aid; decisions on synonymy are left to the user. An initial manual inspection of the top candidates can also easily reveal common synonyms that can be quickly merged at this initial analysis stage.

The result of Step 1 is a list of a wide array of entities that are related to water in the texts of the document corpus. This will typically include water management agencies, but will also include a large number of other organizations, as well as other entities, such as individuals, places, legislation, media outlets, inter alia. This is, in part, a reflection of the way in which our algorithm reduces the list of possible entities, which, as noted above, is skewed to allow false positives. However, it is also potentially useful: these other entities are, in fact, related to water, even though they are not themselves water management agencies. Step 1 results can be useful in informing wider questions about water consumers, users, researchers or advocates in a given region.

4.5.2. Step 2: Purge False Positives

A few entities that are in the results of Step 1 are likely to be spurious, as NLP is imperfect. Step 2 allows users to remove these based on manual inspection.

Additionally, other entities that do exist in the real world and are related to water can be removed, depending on the purposes of the analyses being undertaken. For the larger goals of this article, any entity in the results of Step 1 that is not involved in water management is a false positive; other studies may wish to pursue different directions and, thus, could include any of the wider array of entities mentioned in the preceding section.

In our study, removing all entities except water management agencies results in a list that includes a collection of institutions and organizations that are involved in water management, but typically, this includes institutions that operate at a wide array of scales, from local to federal. This gives insight into the multi-layered nature of water governance in our study areas and demonstrates the character of the wider institutional framework within which local decisions must be made.

4.5.3. Step 3: Apply "Water" Keywords

The final step we employ is to require that the entities listed include keywords from the set of water keywords defined above (set ***W***). This step is intended to winnow the total result set from a large number of potential results into a focused number of results of interest to our analysis. One important effect is that local water management institutions tend to have water keywords in their names, while

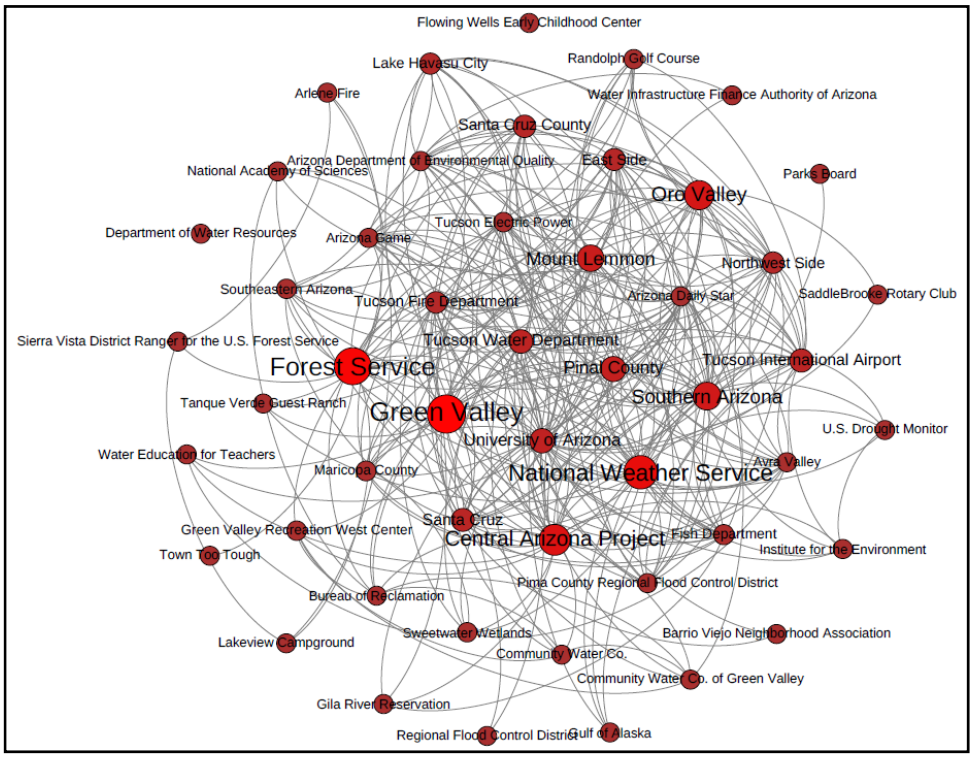
general governmental institutions, especially those at higher levels, such as state or federal agencies, often do not, even when they are clearly involved in water management. The analyst can use this distinction to begin with a very inclusive view, as revealed by Step 1, move to a regional view based on a specific subset of authorities in Step 2 and, finally, move to a more localized view as his or her interests demand. In Step 3 (as in Step 2), some subjectivity is permitted in adjusting these results, so that some entities that have water keywords can be intentionally excluded and others that do not can be explicitly included. The end result, however, should be a reasonable list of entities involved in primarily city-level water management decisions.

The iterative nature and subjective components of the process that we employ here reflect the inherent limitations of NLP and NER and permit flexibility to use parts of our algorithm to investigate questions other than those we pursue here. We note, however, that by allowing manual corrections and, specifically, by allowing individual institutions to be selectively included or excluded, we open the door to subjectivity. For the analyses in this article, we provide documentation for all manual corrections applied to each step below, in the Supplemental Materials. As these network analyses are applied to other regions and in new study areas, we believe we will be able to increase the overall level of automation by reducing and simplifying the degree of manual corrections.

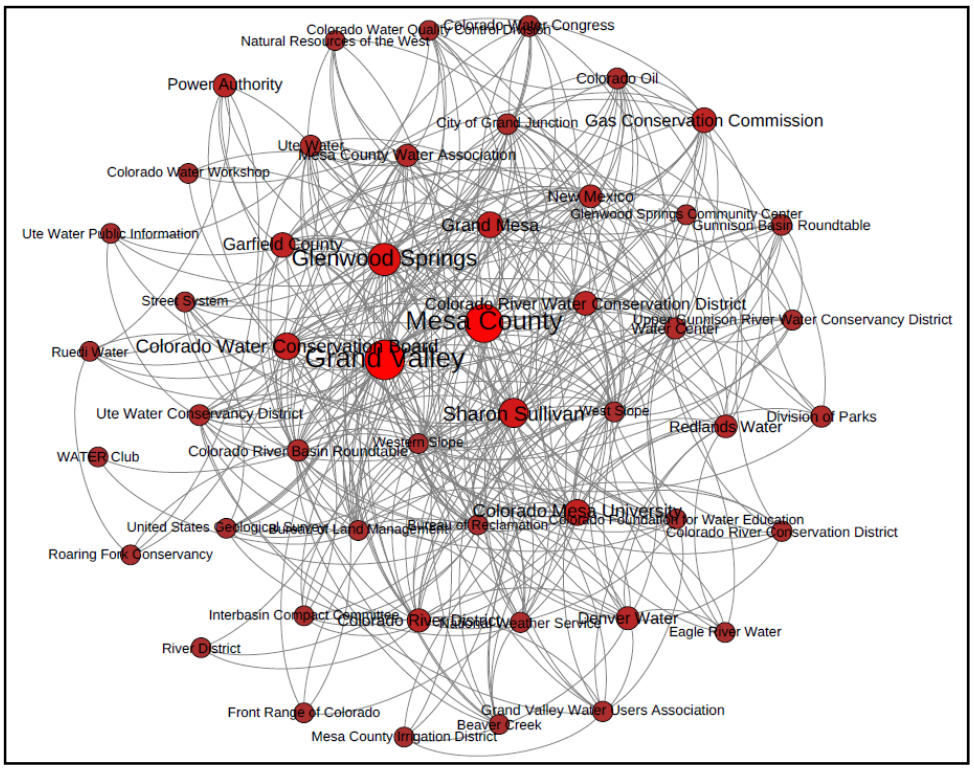
4.6. Network Analyses

Once the list of institutions is established, a network can be constructed among them consisting of the institutions as nodes and edges that link them. Two nodes are connected by an edge if the institutions represented by the nodes appear together in at least one document. We further weight each edge by the sum of the TF scores of the documents in which the two nodes co-occur. These networks can be created based on the results from each of our three steps in our analysis. We note that these networks are most easily apprehended as network diagrams rather than lists of nodes and edges. Network diagrams place nodes onto a page and connect them with lines that represent edges. Different algorithms exist for creating these diagrams. For the diagrams given in Figure 3, Gephi [38,39] is used. For diagrams in Figures 4 and 5, in which types of nodes are differentiated, Cytoscape [40] is used. All variants of all networks are available in the supplemental materials. Each software package includes an array of layout algorithms. For our Gephi graphs, we chose two: the Yifan Hu [41] and Fruchterman–Reingold [42] layouts. The Fruchterman–Reingold layout results in a generally circular arrangement of nodes that dulls the differences among the graphs, but is more legible; versions of the Fruchterman–Reingold graphs are included within the main text for convenience. The Yifan Hu layout conveys the differences among the graphs more effectively, but makes the details less clear on a printed page. In the supplemental materials, we provide both kinds of graphs. Note that in both versions, Gephi calculates a network statistic, betweenness centrality (a measure for each node of the number of shortest-paths from other pairs of nodes that pass through it [39]) for each node and uses this to scale the size of each node and its label. This is echoed in the Cytoscape graphs, but they additionally represent the entity scale (federal, state or local) and the type by shape and color. For the Cytoscape graphs, the "Edge Weight Spring Embedded" layout was used, applied once per graph, using the sum of TF scores for weighs.

Figure 3. Networks generated from Step 1. (**a**) Tucson; (**b**) The Grand Valley; (**c**) Flagstaff; (**d**) Las Vegas.

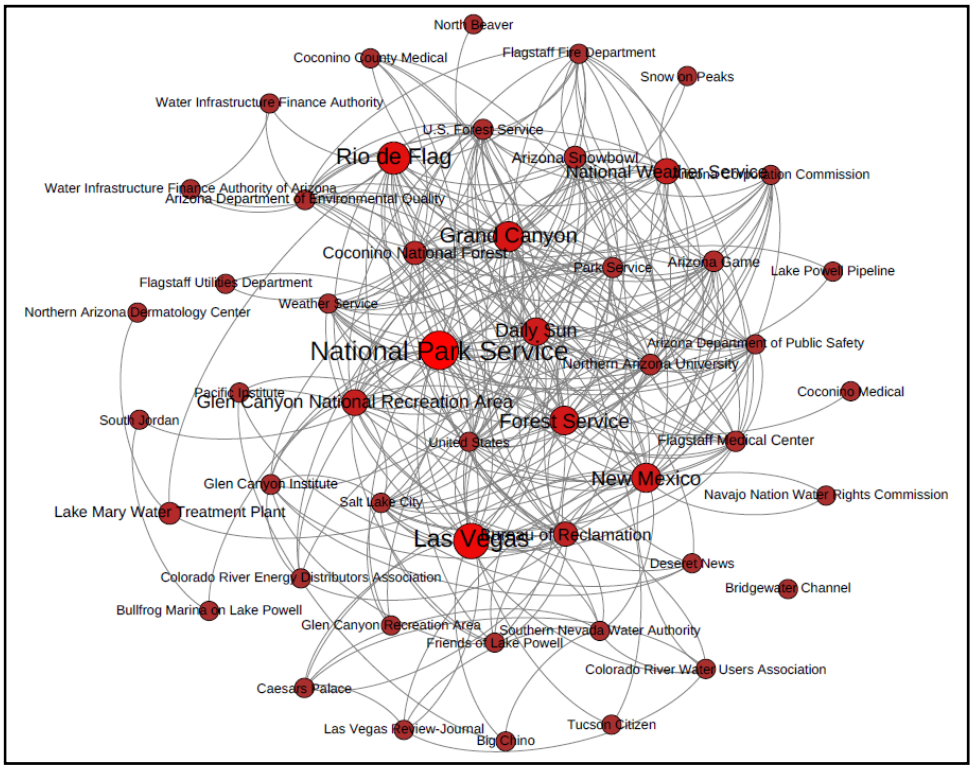


(**a**)

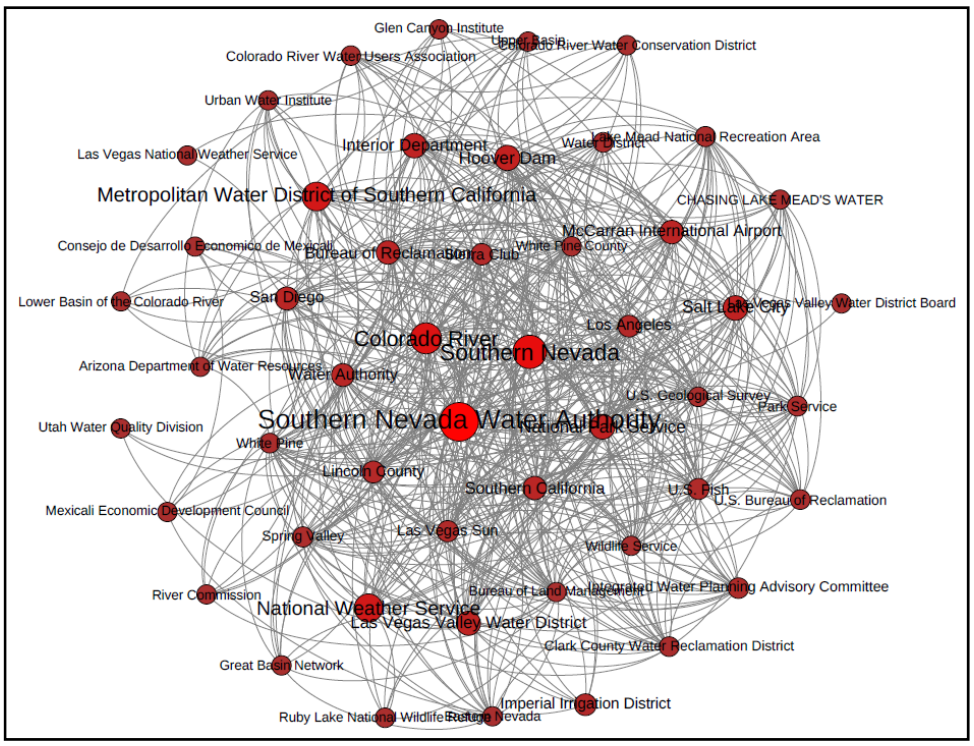


(**b**)

Figure 3. *Cont.*

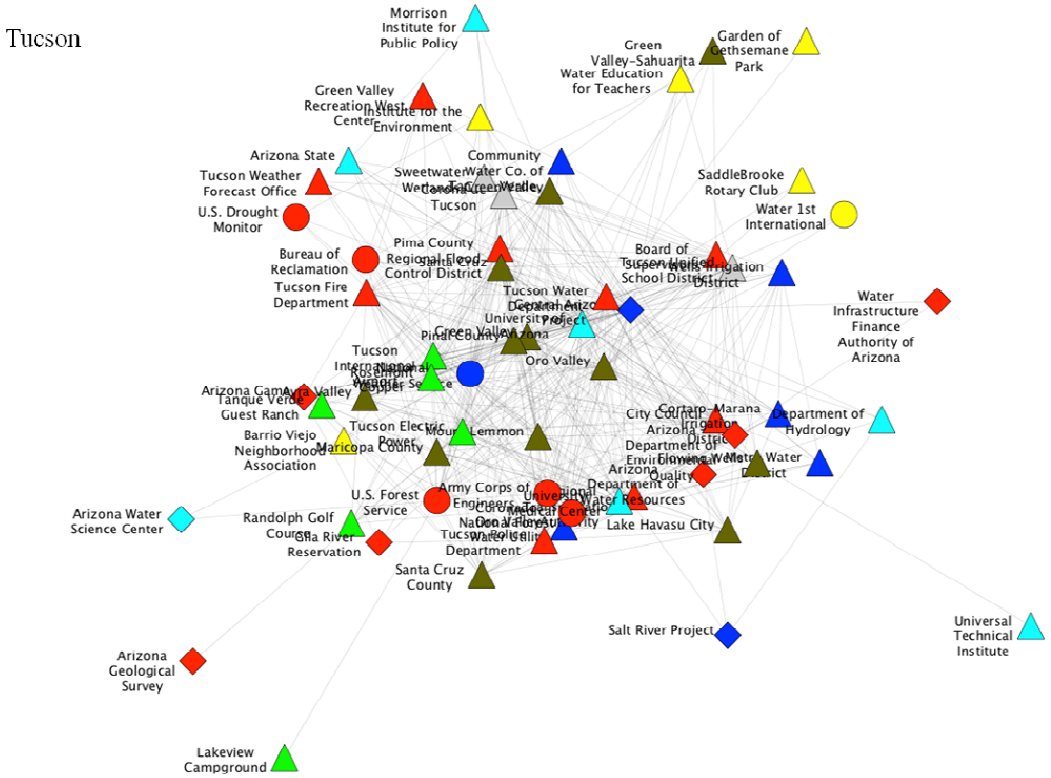


(**c**)

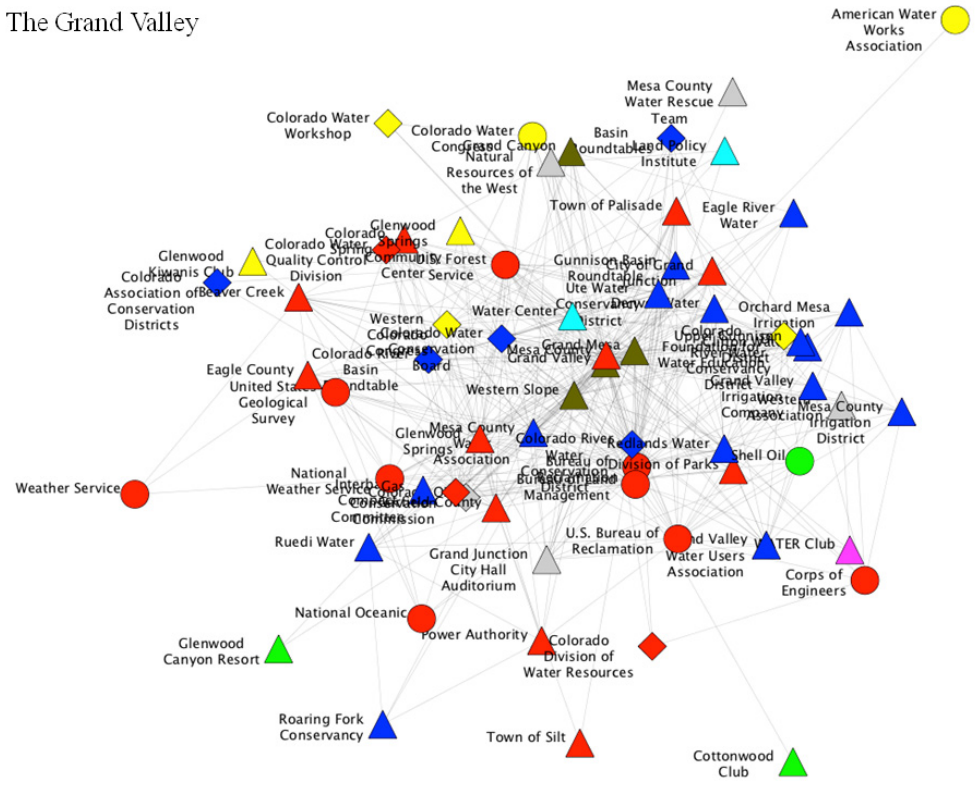


(**d**)

Figure 4. Networks generated from Step 2 results. (**a**) Tucson; (**b**) The Grand Valley; (**c**) Flagstaff; (**d**) Las Vegas.

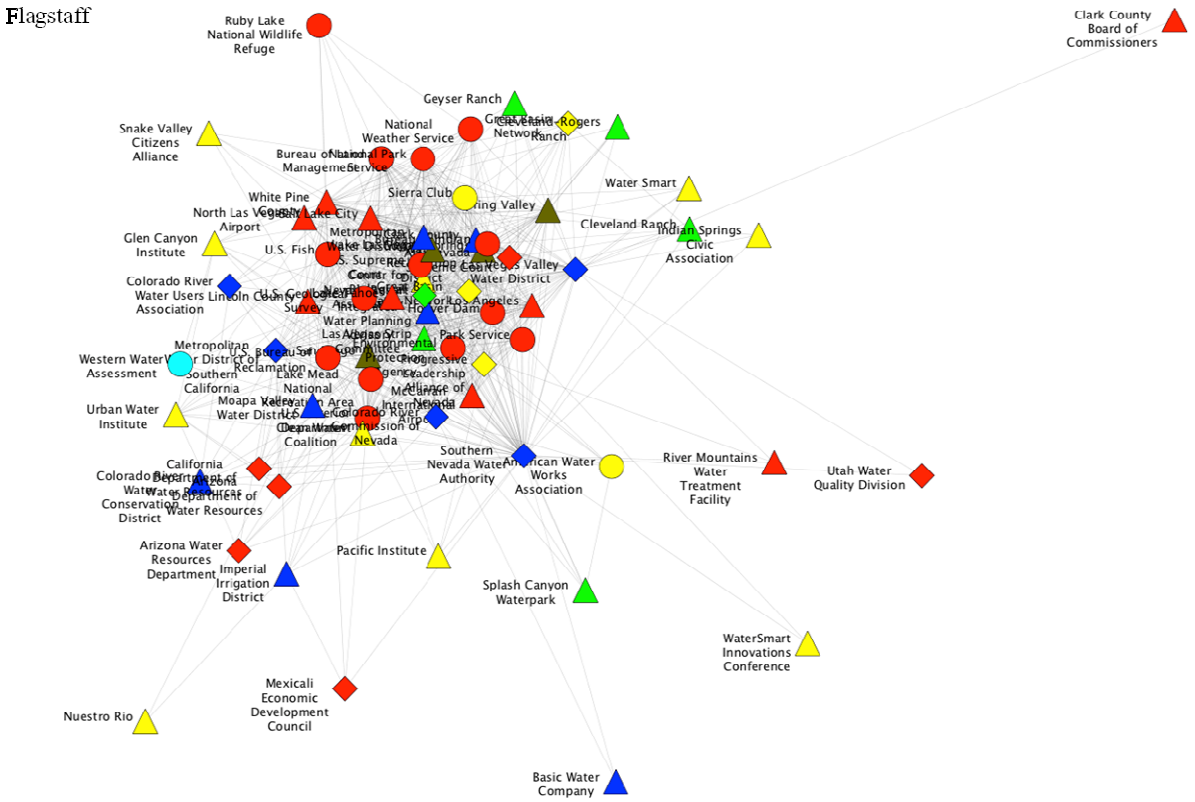


(**a**)

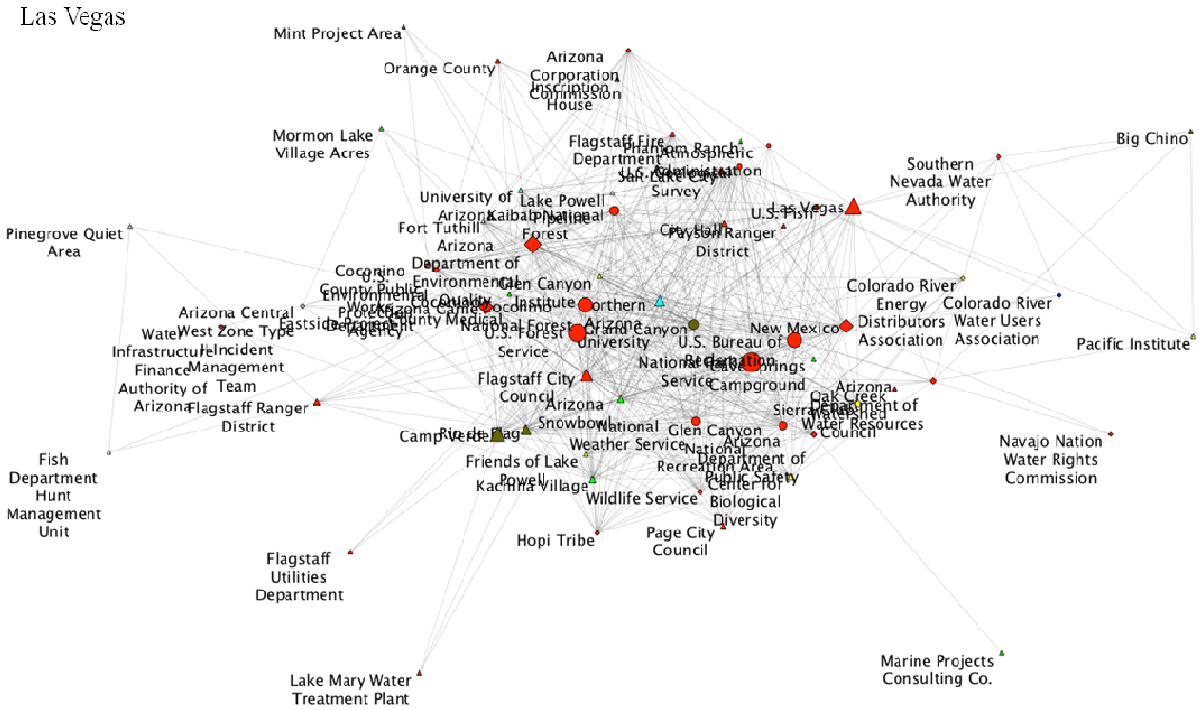


(**b**)

Figure 4. *Cont.*

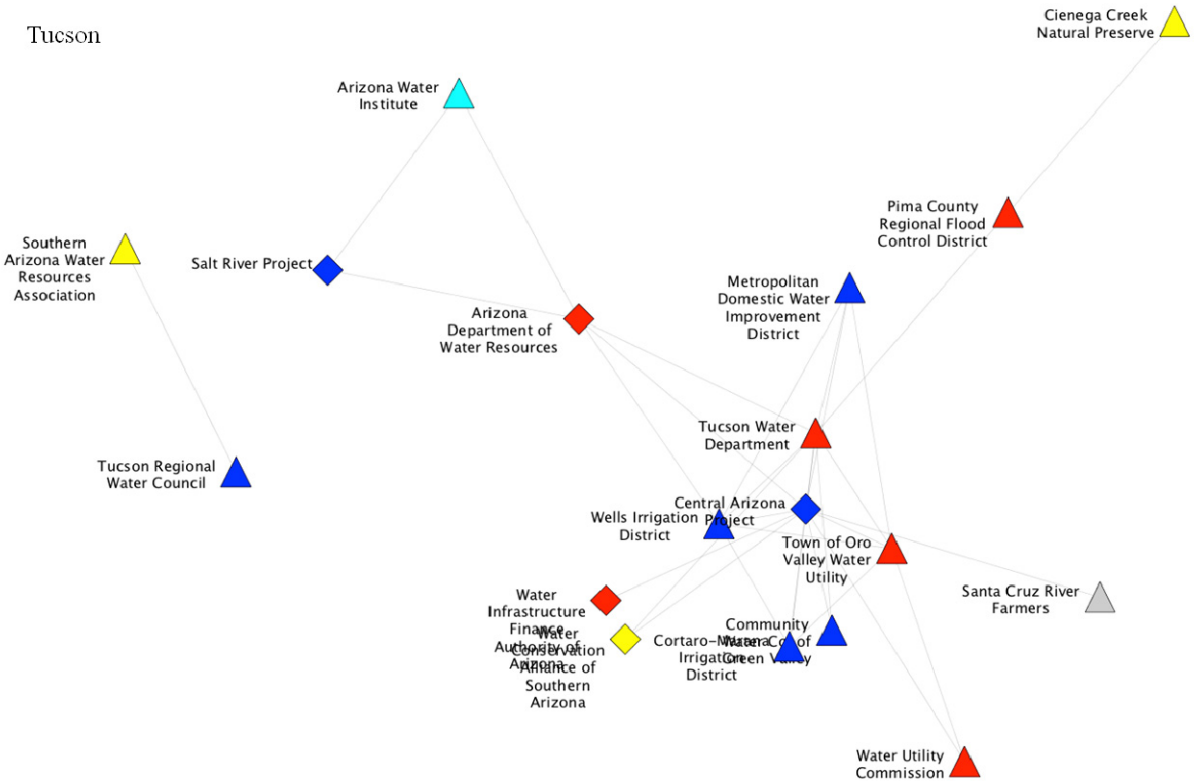


(**c**)

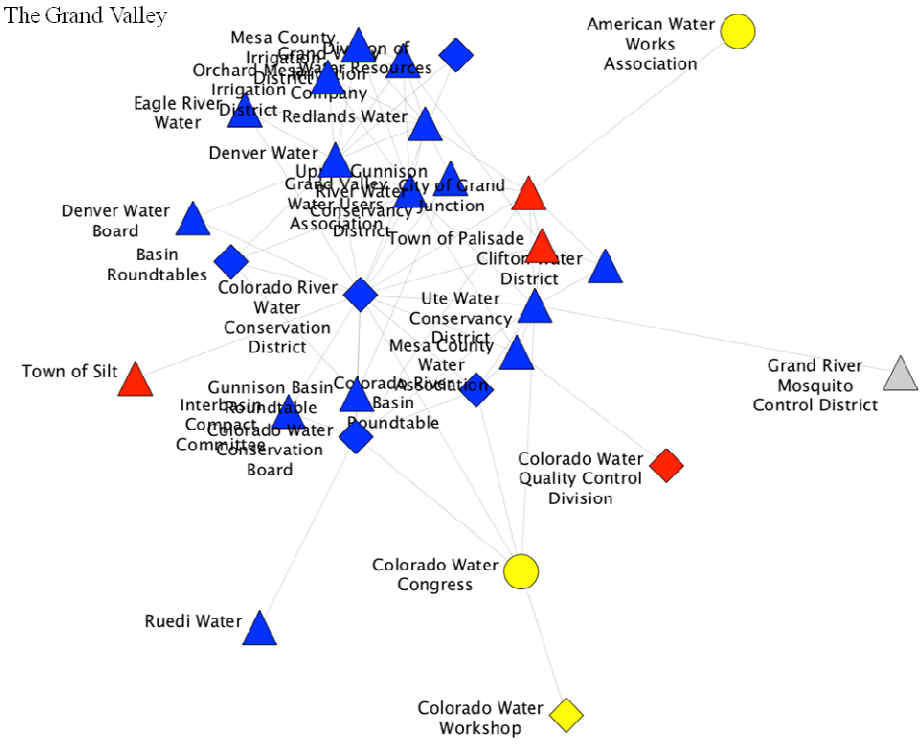


(**d**)

Figure 5. Networks generated from Step 3 results. (**a**) Tucson; (**b**) The Grand Valley; (**c**) Flagstaff; (**d**) Las Vegas.

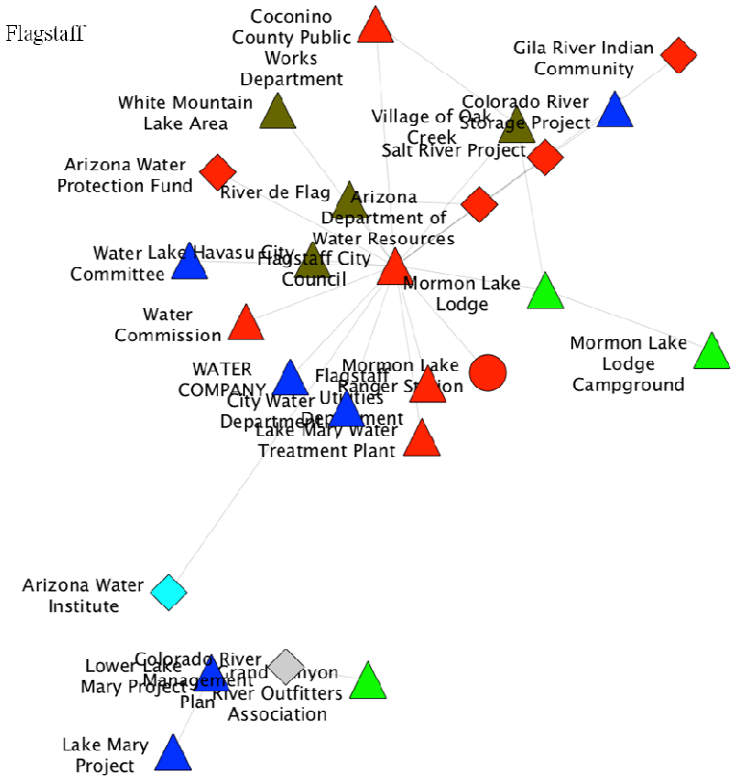


(**a**)

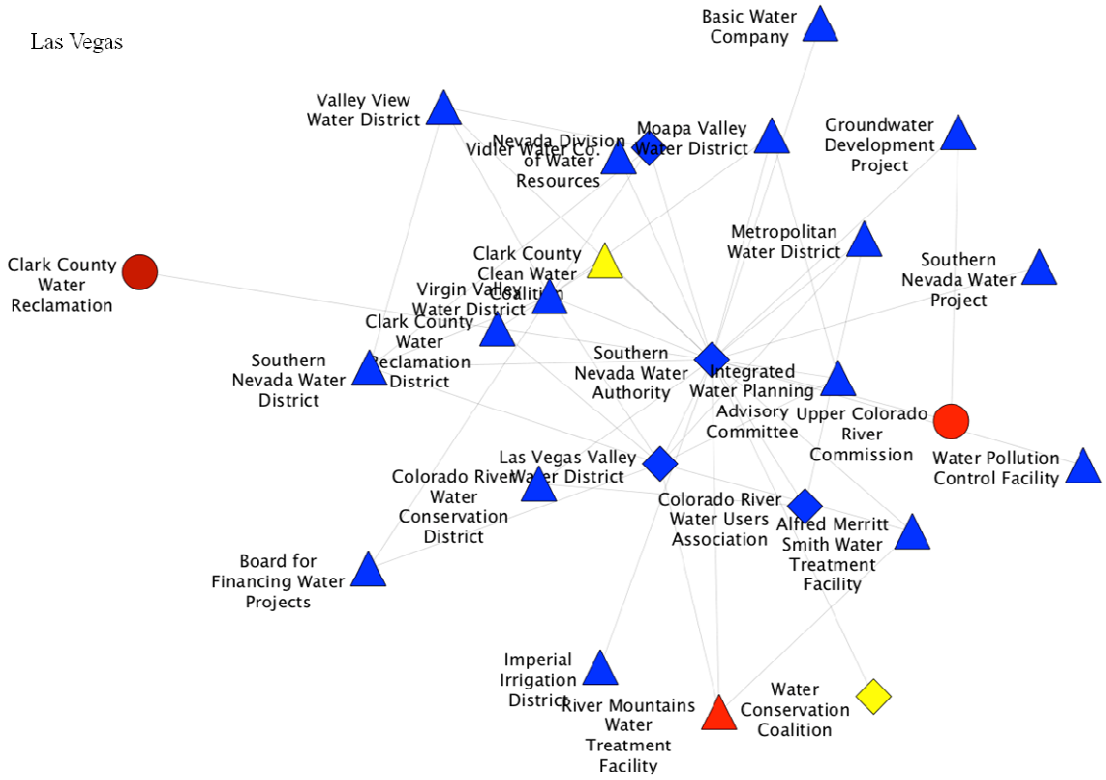


(**b**)

Figure 5. *Cont.*



(**c**)



(**d**)

Quantitative analyses, drawing more generally from social network analysis [43], can be performed on any network thus created. For the networks that result from the earlier steps in the analysis process (*i.e.*, Step 2), betweenness, as defined above, is used; this reflects our interest in determining which nodes in those networks are more central. For our final analysis of Step 3, a different statistic, graph strength [44], is used; graph strength addresses the degree to which a given network is dominated by a centrally-placed node. The number of nodes to which any given node is connected is termed that node's degree. The distribution of degrees is an indication of how connectivity is distributed throughout a network. Graph strength expands on this: instead of a simple count of the edges for each node, the count used is the sum of the weight values that have been assigned to each of a node's edges. A network that has multiple nodes that are connected among themselves will have a distribution of graph strength scores that has multiple nodes with high scores, in contrast to a centralized network, which will have one or a small number of nodes with high scores and the remainder with low scores.

5. Results

The textual data used for this exploratory work was amassed via the process described in Section 4.1 from four regional news sources. The data cover 6–8 years of local news stories; Table 3 gives the sources, date ranges and the numbers of articles collected.

The analysis proceeds from these collections according to the methods described above. We present our results using a strategy that emphasizes the different aspects of our multi-stage process. In the Supplemental Materials, we provide, for each portion of the analysis, a list of the entities extracted and a graphic depiction of the network created using document co-location on all or the top N subset of those entities. However, in the main text, we emphasize specific aspects of each step and present only shortened versions or selected portions of the full data from the supplemental materials. We additionally present summary counts of the manual adjustments made at each step of the analysis, but provide the full lists of such adjustments in the Supplemental Materials. We conclude our results with graphs of the quantitative network analyses performed and make the full tables of values available in the Supplemental Materials.

Table 3. Data sources and articles recovered.

Urban Area	News Source	Time Period(s)	Articles
Flagstaff	*Arizona Daily Sun*	25 August 2005–7 October 2012	11,519
Tucson	*Arizona Daily Star*	1 January 2004–31 December 2011	20,190
Las Vegas	*Las Vegas Sun*	1 January 2005–3 February 2009; 1 January 2011–4 October 2012	66,672
Grand Valley	*Grand Junction Free Press*	13 April 2005–5 October 2012	12,187

5.1. Document Relevance

The supplementary materials give the titles of the top 50 articles ranked by relevance according to our TF scoring. To ensure that the documents recovered are relevant, since it is not always possible to judge the relevance of an article from its title alone, we performed a check by selecting two sample

sets of articles for each of our four data sources. The first set comprised the top 50 articles by TF score; the second was 50 articles randomly selected from the entire corpus. These were manually scored by two independent scorers for whether they should be considered "water related". Some articles were removed from the analysis, because they could not be matched back to unique article scores (generally due to duplicate titles in the original corpus); others were removed due to scorer error (omitted score). The final results of the scoring are presented in Table 4. Within each group of data, the results are shown by whether both scorers rated the article as non-water related (− −), whether they disagreed (− +) or whether they both rated the article as water-related (+ +). Within the top 50 set, the articles are presented in groups representing approximately the top, middle and bottom thirds of the sample by score. Within the random set, the articles are divided into those that had a "zero" TF score (the majority) and then into two approximately equal groups representing the top and bottom halves of the non-zero scores. Kappa scores for inter-observer agreement [45] were 0.699 for the top 50 and 0.545 for the random data sets (using the pooled marginal probability of agreement). The results show that the large majority of top 50 articles are genuinely water-related (few false positives) and that few of the random articles (all of which had lower scores than any of the top 50 articles, for all data sets) were considered water-related (few false negatives). This provides us with the confidence that the methods presented here are robust and sound.

Table 4. Validation results for article relevance scoring.

Scorer Rating		Top 50			Random		
		Top 1/3	Mid 1/3	Bottom 1/3	Top 1/2	Bottom 1/2	Zero
Flagstaff	− −	1	7	3	4	2	33
	− +	1	2		2	3	2
	+ +	13	6	12			
Tucson	− −		2	1	4	3	27
	− +	2	1	1			
	+ +	14	12	13	1	1	
Las Vegas	− −		1		4	4	39
	− +		3				
	+ +	17	13	16			2
The Grand Valley	− −		1		3	4	39
	− +		1				2
	+ +	16	14	16	2		

*5.2. Reduction of Candidate Entities: **C** → **C'** → **E***

Table 5 depicts the course of the analysis in summary form. In general, each step of the analysis process reduces the number of entities under consideration (though it is possible for a step to expand the number of entities, e.g., by synonymizing two examples that each appear in only one document). The table gives the initial number of entities parsed from each corpus and then the number that is parsed from set ***C*** into sets ***C'*** and then ***E***. For manual Steps 1–3, it shows the number and kind of manual adjustments made and the number of entities resulting.

An approximation of the false positive and false negative rate of the movement from ***C'*** to ***E*** is given in Table 6. As part of the development of the algorithm, a subset of articles from each data set

were parsed to create a set equivalent to set ***C***. This resulted in 5521 candidate entities with all scores for internal and contextual attributes. These were hand-coded by a single scorer for whether they represented likely institutions of interest. The equivalent extraction from ***C'*** to ***E*** was then performed and the rates of true *vs*. false positives and negatives calculated. Table 6 represents this by showing four values: first, the total number of entities tested (N); second, the number rated by human scorers to be positives, *i.e.*, the number that should be included in ***E*** (Ind +,); third, the percentage of these that the algorithm correctly identified as positives (% + correct); finally the ratio of false positives to true positives, which reflects how many incorrect entities are placed in ***E*** for every correct one (false + fraction). Note again that our algorithm is biased toward false positives.

Table 5. Numbers of entities and removals at each stage of the analysis processes.

Stage	Flagstaff	Tucson	Las Vegas	Grand Valley
Total Distinct Entities (*C*)	59,914	92,874	271,824	63,684
Total Distinct Entities in > 2 Documents (*C'*)	6,309	6,668	24,514	7,152
Entities Meeting Institution Criteria (*E*)	2,709	2,681	8,864	3,037
Step 1:				
Number of Synonym Corrections	11	11	9	5
Resulting Number of Entities	2,706	2,678	8,863	3,038
Step 2: (Top 100 of list from Step 1)				
Synonym Corrections	11	6	13	9
False Positives Removed	27	32	20	25
Resulting Entities (Of Top 100)	63	64	67	66
Step 3:				
Synonym Corrections	1	4	10	1
Water Keywords Manually Set to "True"	4	2	0	7
Water Keywords Manually Set to "False"	75	35	180	42
Final Number of Entities	42	33	32	35

Table 6. Estimate of ***C'*** -> ***E*** effectiveness.

Data Source	N	Ind +	% + Correct	False + Fraction
Arizona Daily Sun (Flagstaff)	1518	342	92%	1.42
Arizona Daily Star (Tucson)	1014	226	92%	1.02
Las Vegas Sun	952	158	96%	1.81
Grand Junction Free Press	2037	453	83%	1.45
Overall	5521	1179	89%	1.41

The results of our extraction of water authorities for all four data sources are shown in Table 7; only the top 20 results are shown, with the total number of authorities with non-zero values for relevance to water topics given, as well. The Supplementary Materials include the top 50 results.

Above, we presented a brief summary of the rates of false positives and false negatives that our algorithm produces; a somewhat different question is whether the list of authorities generated matches the actual collection of entities in the real world; that is, does our list include all the water management authorities that are actually found in our study areas?

Table 7. Top 20 water-related institutions from the basic extraction algorithm.

Flagstaff	**Tucson**
1. Caesars Palace	1. National Weather Service
2. Colorado River Water Users Association	2. Tucson Water Department
3. Southern Nevada Water Authority	3. Tucson International Airport
4. National Park Service	4. Pima County Regional Flood Control District
5. New Mexico	5. Central Arizona Project
6. North Beaver	6. Tucson Fire Department
7. Colorado River Energy Distributors Association	7. Mount Lemmon
8. U.S. Forest Service	8. Arizona Daily Star
9. Bureau of Reclamation	9. University of Arizona
10. Las Vegas Review-Journal	10. Green Valley
11. Daily Sun	11. Tanque Verde Guest Ranch
12. Northern Arizona Dermatology Center	12. Town Too Tough
13. National Weather Service	13. Southern Arizona
14. Forest Service	14. Randolph Golf Course
15. Snow on Peaks	15. Sweetwater Wetlands
16. Glen Canyon National Recreation Area	16. Arlene Fire
17. Deseret News	17. Lakeview Campground
18. Big Chino	18. Sierra Vista District Ranger for the U.S. Forest Service
19. Grand Canyon	19. Oro Valley
20. Las Vegas	20. Northwest Side
Las Vegas	**Grand Junction**
1. Lower Basin of the Colorado River	1. Water Center
2. Bureau of Reclamation	2. Colorado Mesa University
3. Colorado River	3. Upper Gunnison River Water Conservancy District
4. Southern Nevada Water Authority	4. WATER Club
5. CHASING LAKE MEAD'S WATER	5. Mesa County
6. Southern Nevada	6. Colorado Water Conservation Board
7. Water Authority	7. Redlands Water
8. Interior Department	8. Interbasin Compact Committee
9. Lake Mead National Recreation Area	9. Gunnison Basin Roundtable
10. River Commission	10. Colorado River Basin Roundtable
11. Hoover Dam	11. Bureau of Reclamation
12. National Park Service	12. Grand Valley Water Users Association
13. White Pine	13. Grand Mesa
14. Colorado River Water Users Association	14. Glenwood Springs
15. Urban Water Institute	15. Grand Valley
16. Arizona Department of Water Resources	16. River District
17. Metropolitan Water District of Southern California	17. Glenwood Springs Community Center
18. Southern California	18. Colorado River District
19. Las Vegas Valley Water District	19. Mesa County Irrigation District
20. Upper Basin	20. Colorado Foundation for Water Education

This is an important, but difficult question. One of the advantages to our procedure is that it reveals the involvement in water issues of institutions or organizations that might not otherwise be thought to

play a role. Media outlets, research institutions, non-governmental organizations, clubs, commercial entities and other kinds of organizations can be seen to be associated with discussions of water, and their roles in water management explored more fully on the basis of this insight. The reverse of this—independently compiling an exhaustive list of organizations involved in water discussions—is much more difficult.

Nevertheless, it is reasonable to compile a list of water-related institutions through some other means and then to ask whether these organizations appear in our results. This will provide one approach to validate our method. What should be noted, however, is that there are two markedly different ways that an entity on this list could fail to appear in our analyses: one is that our method fails to capture the entity from within our data source, because the technique used fails to recognize it or classify it correctly; the other is that the entity is never mentioned in our source data at all.

To investigate this, we selected one of our study areas (Tucson) and compiled a list of water utilities from sources outside our corpus [46–48]. This focused specifically on water utilities and, therefore, ignores many other kinds of entities that we might capture in our analysis; it is also a limited sample that makes no claim of being exhaustive. The source data and NLP analysis were examined to determine whether the name of the entity appeared in the data set and, if so, whether it was captured by our analysis or at what point it was missed or culled. The points at which a given entity could be eliminated were: because the name was not recognized as a candidate authority and, thus, never placed in set **C**; because the phrase did not appear in more than one document, and so was not placed in **C'**; or because the phrase did not meet the criteria required to be considered a water authority and, thus, was not placed in set **E**. The list of terms searched and the result for each is provided in the Supplementary Materials. The results indicate that the large majority of institutions that were absent from our analysis were absent because they did not occur in the corpus. Of 75 unique institutions mentioned by our outside sources, 54 could not be found in the original corpus of newspaper text; of the remaining 21, eight are omitted because they occur in too few valid documents; the remaining 13 persist through the remainder of the analysis and are present in ***E***.

In general, however, these omitted entities are typically the "small-scale" providers that fall below our level of interest. The Supplemental Materials additionally give the size, in number of residences served, of many of the water providers from our outside sources; the figures are taken from a NY Times online resource [49]. An exact assessment is made difficult by the fact that entities can be referred to by different names, and sizes are available for only a portion of the elements in our test; however, it is possible to use these data, even if incomplete, to gain a sense of what may be missing from the overall analysis. One provider serving over 10,000 people, another serving 8000 people and nine serving between 1000 and 6000 were missing from the corpus; 16 other providers serving under 1000 people were also missing, yielding a total of about 57,000 people. The NY Times data set includes about half of the entities in the original list; if it is assumed that the other half are essentially a matching set (a very conservative assumption), then it is possible to conclude that water providers serving about 114,000 people were never mentioned in the newspaper corpus. Similarly, five providers in the NY Times data set appeared in the corpus but were dropped by the analysis (although the largest may appear (and be captured) under an abbreviated name); these five include about 26,000 people, which can be adjusted proportionally to include the additional three not mentioned in the NY Times

data set, and arriving at about 42,000 people. In total, it can be estimated that that there are about 156,000 people whose water providers do not appear in our analysis.

While this large number is not inconsiderable, it is also not discouraging with respect to our purposes. First, the estimates to fill in the missing providers (not in the NY Times data) and to deal with the other ambiguities between our sources and the NY Times values are likely to be high. Second, the bulk of this number comes from a range of providers with only two serving over 10,000 people, a few with between 5000 and 10,000 people and the rest with numbers in the low thousands or hundreds. Third, and perhaps most importantly, this must all be considered against the sizes of the providers we did capture: Tucson Water, according to the same NY Times data set, serves more than 650,000 people; two other providers serve more than 40,000 each, and all of the providers for which we have matching figures from the NY Times data set serve nearly 790,000 people. The boundary at which a provider may not be captured may be around 15,000 people (about the size of the smallest provider captured and the largest one missed), which is comparatively small in an area with a population approaching one million people, and many of the providers missed serve under 1000 people.

5.3. Step 1: Remove Synonyms

The procedure to remove synonyms for initial inspection involves simple pattern-matching: common abbreviations (e.g., "City Water Company" *vs*. "City Water Co.") can be easily recovered. Figure 3 shows graphs of the networks that can be created using document co-location for the top 50 elements that remain in the list of entities after these synonyms are resolved. Note that the top 20 are unchanged from the preceding step for all four data sets, as given in Table 7.

The expectation for this stage of our method is that the lists of entities will include a wide variety of kinds of entities, from federal, state and local institutions to media outlets, universities and even individuals, along with some pure false positives (spurious or improperly parsed elements). This is true even though we attempt to filter out many of these categories explicitly. The lists given in Table 8, the graphs in Figure 3 and the full lists in the Supplementary Materials bear this out. A categorization of the elements in the top 50 of each source area is given in Table 8, with the full table and coding employed given in the Supplemental Materials. Of special note is the very small number of entries that are persons: personal names are very effectively eliminated by our procedure.

5.4. Step 2: Remove False Positives

For Step 2, the top 100 elements in each list were considered; these were inspected, and false positives and newly recognized synonyms were removed. The limitation to a number of nodes on the order of 50 is still desirable; this seems to capture a wide range of entities, but result in a useful graph. The strategy employed was to reduce the top 100 elements and allow the list of elements to shrink (instead of adding elements further down in the list to replace those removed). The full lists, and the list of removals and synonymizations, are given in the Supplementary Materials. Figure 4 presents graphs of the networks created from the resulting entities.

This step deliberately includes federal, state and regional agencies, but eliminates as false positives entities that fall into the other categories given in Table 8 It can be considered a manual extension of the formula for moving from $\boldsymbol{C'}$ to $\boldsymbol{E}$ given above, improving on the automated procedure.

Table 8. Counts of entities by category in the Top 50 after Step 1 of the analysis.

Category	Flagstaff	Tucson	Las Vegas	Grand Junction
Act		1		
Business	5	4	6	5
Event	2	4	2	2
Local Government (Non-Water)	4	6		4
Local NGO	2	1	1	1
Media	5	1	1	2
National NGO	1	1	1	1
Person				4
Place	27	21	25	18
Research Org.	3	9	2	6
Road		3		
Federal (Water)	15	7	16	11
State (Water)	8	8	5	5
Regional (Water)	9	3	15	12
Local (Water)	6	18	12	16
Other/Spurious	13	13	14	13

Note that Figures 4 and 5 (as well as the Cytoscape graphs of the Step 1 results, which are presented in the Supplemental Materials for reference) represent entity scale and type according to the following conventions. Scale is represented by node shape as: federal = circle; state = diamond; and local = triangle. Colors and entity types correspond to: blue = water management-specific institution; red = government agency; green = business; yellow = non-governmental organization; cyan = university; magenta = media; aqua = person; brown = place; gray = unknown.

5.5. Step 3: Water Keywords

When each of the study areas' lists are further restricted to only entities that include words from ***W***, they become manageable in their entirety; thus, instead of limited analysis to the top 50 or 100 entities, as above, we can review each entire list in a way that was not possible at the end of Step 1 (when, for example, Las Vegas's list of entities included more than 8000 entries; with water keywords applied, the list is reduced to about two hundred). Using the whole list means removing a fairly large number of false positives, but the resulting lists include a very small numbers of nodes.

For each study area, the list of special inclusions (terms that actually contain no water keyword, but that we explicitly bring into this final step), the list of false positive removals and any final synonyms is given in the Supplemental Materials. Figure 5 depicts the graphs generated from these lists.

5.6. Network Analyses

The graphic depictions of the networks suggest differences among the four study areas; these differences can be elucidated further using quantitative network statistics.

Certain analyses can be performed for the results from Steps 1 and 2. Step 1 results, however, include many spurious elements and further analysis is omitted in favor of considering Step 2. Before doing so, it should be noted that Step 2 results are also somewhat provisional; the NLP process

presented here is concerned mainly with eliminating categories of entities by the end of Step 3, and the earlier results are incomplete and may be biased toward certain entities in unexamined ways (possibly based on real-world characteristics, but also possibly based on arbitrary textual variation). However, the collection of entities revealed is still of interest, in that it reflects a wide constellation of water-related entities. Table 9 shows a list of the counts of entities categorized along two axes: scale and type of institution. Scale is one of three values, roughly corresponding to federal, state and local; the types of institutions are those listed above as differentiated in the Cytoscape graphs. Note that the categorization offered is useful, but is not formalized. In some cases, there is an ambiguity in either scope (e.g., a state institution from Arizona appearing in Las Vegas's data set might be considered to be operating at a level higher than state) or in type (especially "place", because place names can also represent municipal institutions, but also among government *vs*. non-government institutions). The results here are offered to suggest possible future work in which the Step 2 results are refined and achieve higher confidence.

Table 9. Counts of entities, by category and level, from Step 2 results.

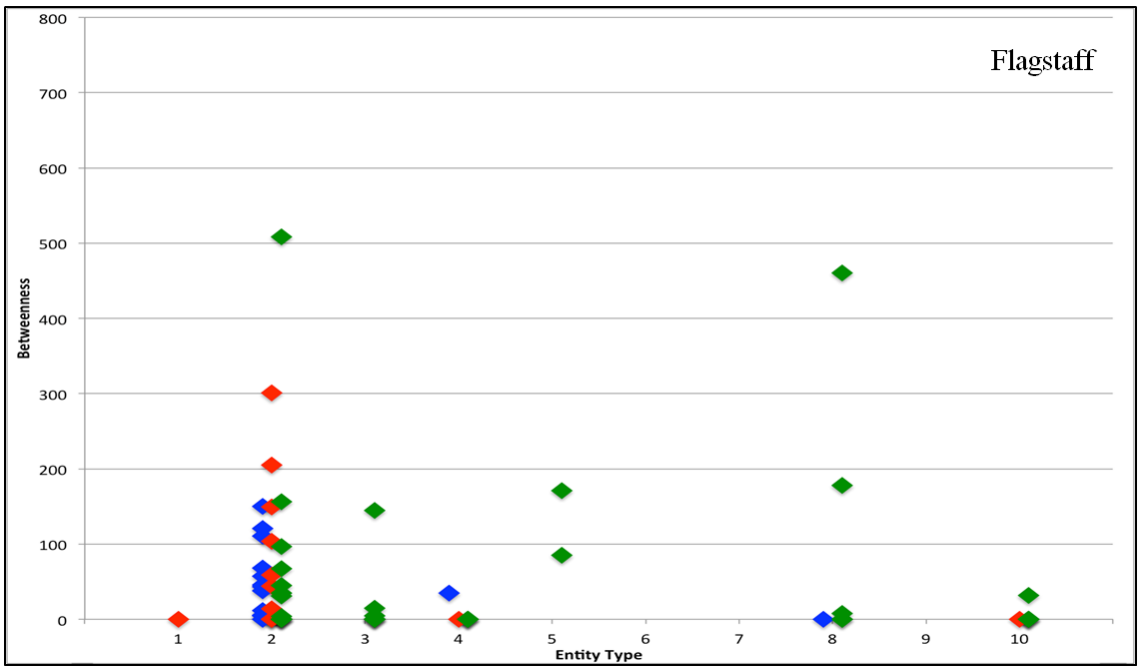
Source	Scope	Water 1	Government 2	Business 3	NGO 4	University 5	Place 8	Other 10	Total
Arizona Daily Sun	Federal		12		1		1		14
	State	1	10		1			1	13
	Local		13	7	4	2	5	5	36
	Total	1	35	7	6	2	6	6	63
Arizona Daily Star	Federal	1	5		1				7
	State	2	6			1			9
	Local	5	9	6	5	6	12	3	46
	Total	8	20	6	6	7	12	3	62
Las Vegas Sun	Federal		14		2	1			17
	State	5	6	1	3				15
	Local	7	9	5	10		4		35
	Total	12	29	6	15	1	4		67
Grand Junction Free Press	Federal		9	1	2				12
	State	5	3		3			1	12
	Local	15	11	2	2	2	4	5	41
	Total	20	23	3	7	2	4	6	65

Table 9 gives the raw counts for each of the four study areas of Step 2 entities by level and type. To assess the positioning within the network of each node, betweenness was calculated; this version of the calculation of path length uses the summed TF scores of the edge weights (which makes the values different from the statistic mentioned above that is used to scale the nodes). The distributions of the scores by node, type and level are given for each of the four areas in Figure 6 (using the numeric codes from Table 9). The raw data is provided in the Supplemental Materials.

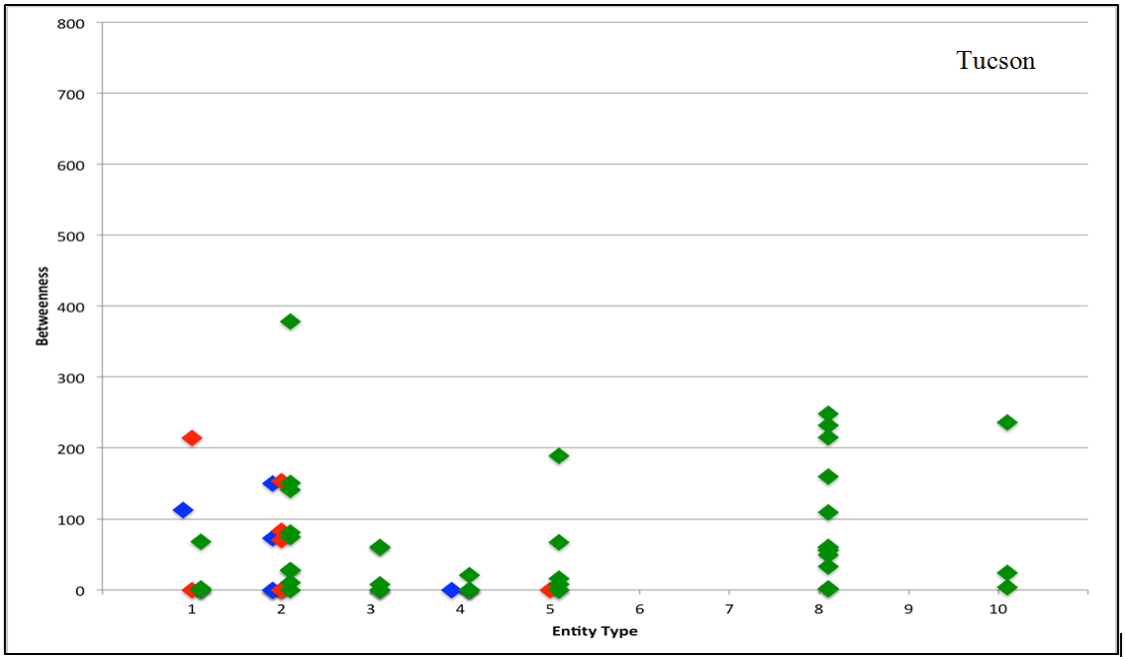
To better quantify the differences among the Step 3 networks, the "graph strength" for each network was calculated. Graph strength is calculated by calculating for each node in the network the sum of the weights of its edges. The weights in this case are the TF scores of the documents used to create the

edges. This weighting further emphasizes water-related documents, allowing us to highlight relationships among entities that are found in water-related discussions. The values that result are given in the Supplemental Materials and are depicted graphically in Figure 7, where the values are normalized and displayed in descending order.

Figure 6. Distributions of betweenness centrality, by type and level of entity, from Step 2 results. Blue = federal; red = state; green = local. Categories 1–10 are as given in the main text. (**a**) Flagstaff; (**b**) Tucson; (**c**) Las Vegas; (**d**) Grand Junction.

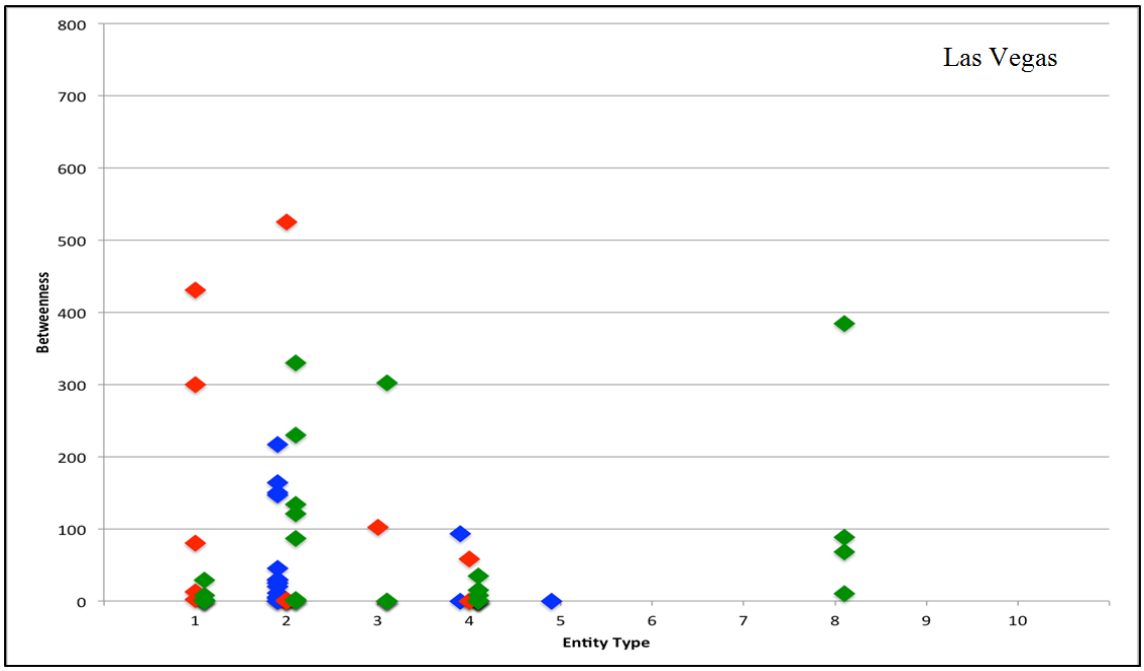


(**a**)

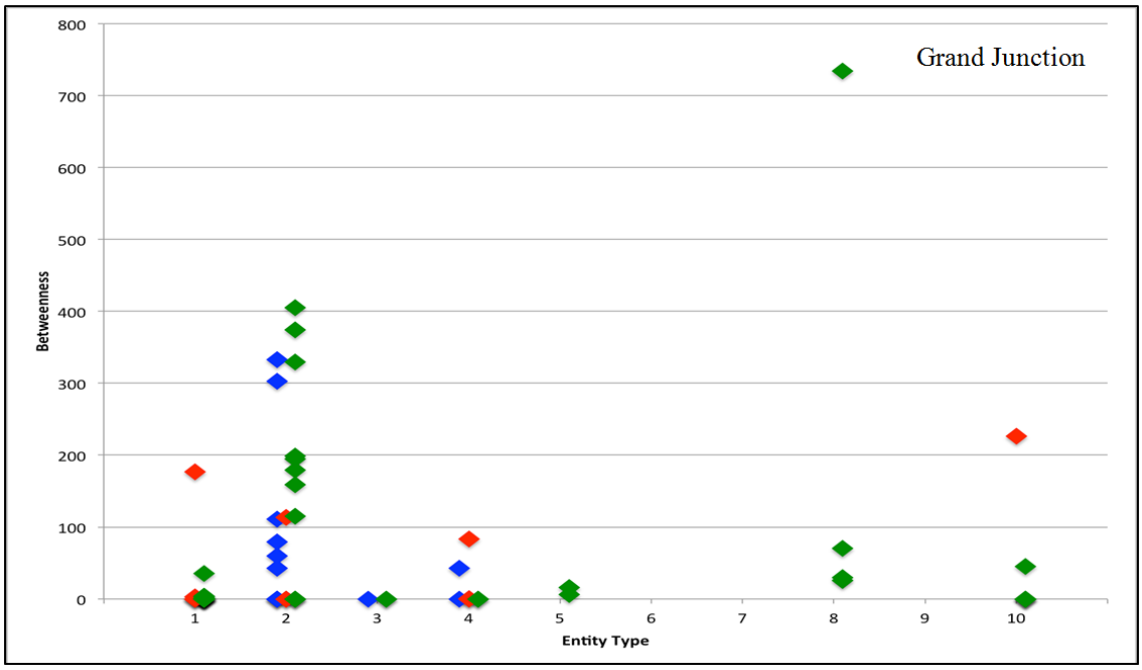


(**b**)

Figure 6. *Cont.*



(**c**)



(**d**)

6. Analysis and Discussion

6.1. Step 1 Analysis

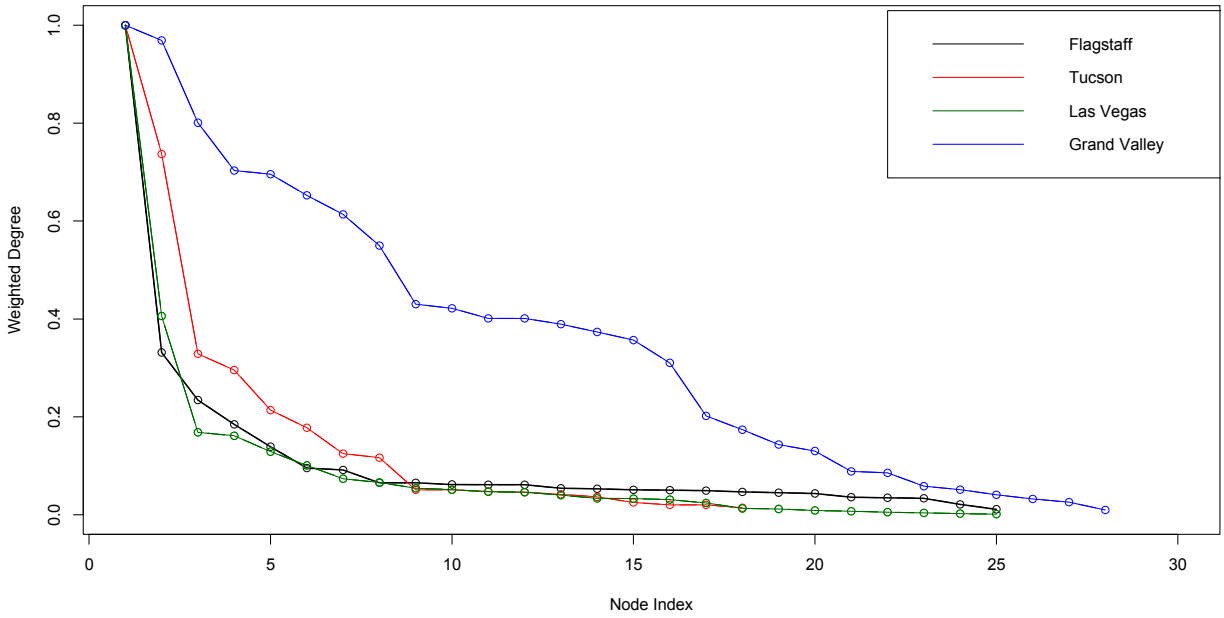
The results yielded by Step 1 of the analysis give a picture that includes a wide range of kinds of institutions; notable are the appearance of media (e.g., Tucson Citizen), institutions related to energy (e.g., Colorado River Energy Distributors Association), research institutes (e.g., Urban Water Institute),

as well as other elements, such as individuals. These institutions and individuals are components of the water infrastructure, and their network of connections might be of use in an array of different analyses. Because our algorithm has attempted to filter some of these out to find only institutions, their appearance in any of our results is artifactual and not exhaustive. However, alternative versions of our algorithm could be created that target these other kinds of elements and examine the networks of connections among them.

6.2. Step 2 Analysis

The results from Step 2 that are given in Table 9 and Figures 4 and 6 include the results of Step 1 filtered to include water management institutions more strictly than Step 1, but not restricted to scope. They thus include federal, state and local institutions, as well as non-governmental organizations operating at all scales. The raw counts of institutions are illuminating: all four data sets are approximately the same size (the number of nodes is between 63 and 67), but Grand Junction has 15 local water management institutions, with Las Vegas in a distant second place with only seven. Tucson has twice as many (12) entities that are categorized as "places" than any of the other three; while it should be mentioned that "place" is a problematic category, because a place name may also be used as an institution, the list includes the names of adjacent counties and, so, may also reflect a relationship with the areas around Tucson proper. Las Vegas has more NGOs than any of the other areas, and Flagstaff has noticeably more government entities (at all levels) than the other areas, although government entities are strongly represented in all areas.

Figure 7. Normalized weighted degree distribution (sum of TF scores on common documents).



The graphic depictions of the Step 2 results in Figure 4 are also illuminating. The Las Vegas data set is clearly different. The image provided in the text has been rescaled; the original (available in the Supplemental Materials) is heavily impacted by the fact that the number of edges is higher and the layout algorithm is less effective at moving central components away from peripheral ones. The central cluster includes a wide mix of mostly governmental institutions, many of them federal, but also NGOs,

businesses and water-specific organizations. The clustering effect of the layout algorithm is reversed in the (unscaled) Flagstaff map, where nodes are spread much more thinly. The centrality of several federal institutions—the National Park Service and two national parks—is clear. For the Tucson data set, the appearance in the central area of Tucson Water, Green Valley, Oro Valley and the Central Arizona Project is expected, while state and federal institutions (Water Infrastructure Finance Authority of Arizona, Bureau of Reclamation) are pushed to the periphery. This is echoed in the Grand Valley data set, but more complicated, because of the existence of numerous local entities and, as will be shown, a different overall network structure.

The centrality of these entities within the detected networks, as shown via the betweenness statistics, whose distribution is given in Figure 6, gives still more information. Flagstaff has only one specifically water-related institution (the "Colorado River Water Users Association"), and it does not occupy a central node. Conversely, numerous government institutions occupy more central nodes; the most central node is "Las Vegas" (here classified as a local government, not a place). Many of the other central nodes are federal, state and local institutions; there are nine such institutions with betweenness scores above 100 and only four other entities of any kind above this threshold.

In our Tucson data set, the most central node is "city council"; this reflects the central position of the Tucson city government. The large number of places has already been noted, and they score highly on the betweenness statistic. The Central Arizona Project is clearly a prominent water management institution; a number of entities associated with the University of Arizona are also prominent, as is the Tucson Unified School District (probably for its role as a consumer of water and as a vehicle for conservation efforts).

The node in the Las Vegas data set with the highest betweenness score is the Nevada Supreme Court. Second is a water-specific entity, the Las Vegas Valley Water District. Of interest is the fact that the Southern Nevada Water Authority is comparatively low, with 17 entities more central than it. All but four of these are government agencies, federal and local, but, except for the Supreme Court, not state. The other prominent entity is San Diego (classified here as a place).

Grand Junction's graph is dominated by a very high-valued datum; the term "western slope", classified here as a place, is the most central node by far. It is likely, however, that this is merely because the name of the region appears in a wide range of articles (unlike other place names, it is unlikely to also represent an institution). Although there are a large number of water-specific entities, most of these do not occupy nodes for which the betweenness statistic indicates a high value. Conversely, 12 government agencies have scores above 100, along with only the Colorado River Water Conservation District and a difficult-to-classify "Colorado Oil".

6.3. Step 3 Analysis

Our specific interest is in discerning systems in which water management is distributed among numerous peer institutions *vs*. those in which water management is dominated by single, central players. Our Step 3 data suggest that the Grand Junction example is a case in which multiple peer institutions interact on a nearly equal basis, while the other three are dominated, albeit to varying degrees, by a single central player. The quantitative analysis using the graph strength statistic on the Step 3 results confirms the Colorado data set's differences. Our expectation is that the Grand Junction

data set will have a large collection of nodes that all have high values for graph strength, indicating that they are all highly connected to immediate neighbor nodes (including each other), but that the data from the other area will have only a small number of highly connected nodes and the rest will have much lower values.

Table 10 gives selected results drawn from Figure 7, grouped into value ranges; the full table of results is available in the Supplementary Materials. The difference between the Grand Junction data set and the others is highlighted in the figure, but to be certain that it is statistically different, a pairwise Wilcoxon rank sum test among the four data sets was performed; the results are given in Table 11. The Wilcoxon rank sum test orders all measurements (from all samples) and asks if the order is random or is biased with respect to one of the sample sets. If one of the sample sets differs significantly from the others (e.g., by having consistently higher values), the test statistic will show this difference, and the null hypothesis that the underlying distribution of both samples is the same can be rejected. Using this test, statistically significant differences (p-value < 0.05) are seen between the Grand Valley and all three other study areas. These results indicate that the Grand Valley has a highly fragmented system: there are more nodes that are more central, and these are more highly connected to each other.

Table 10. Selected values from the graph in Figure 7, grouped by value.

Weighted Degree	Flagstaff	Tucson	Las Vegas	The Grand Valley
0.8–1	Flagstaff City Council	Central Arizona Project	Southern Nevada Water Authority	Colorado River Water Conservation District Ute Water Conservancy District City of Grand Junction
0.6–0.8		Tucson Water Department		Redlands Water Denver Water Colorado River Conservation Board Town of Palisade
0.4–0.6			Las Vegas Valley Water District	Grand Valley Water Users Association Mesa County Water Association Grand Valley Irrigation Company Mesa County Irrigation District Orchard Mesa Irrigation District
0.2–0.4	Arizona Department of Water Resources Salt River Project	Town of Oro Valley Water Utility Wells Irrigation District Cortaro-Marana Irrigation District		Colorado River Basin Roundtable Clifton Water District Upper Gunnison River Water Conservancy District; Gunnison Basin Roundtable; Interbasin Compact Committee

Table 11. Wilcoxon ranked sum test on normalized weighted degree distributions, pairwise among all data sets in Figure 7. Significant test results shown in bold.

Ranked Sum vs.	**Flagstaff** $n = 25$ **median = 0.054**	**Tucson** $n = 18$ **median = 0.050**	**Las Vegas** $n = 25$ **median = 0.040**	**Grand Valley** $n = 28$ **median = 0.365**
Tucson	W = 222.5 $p = 0.9607$			
Las Vegas	W = 404.5 $p = 0.0758$	W = 296.5 $p = 0.0805$		
Grand Valley	**W = 163.5** **$p = 0.0009$**	**W = 143.5** **$p = 0.0151$**	**W = 124.5** **$p < 0.0001$**	

This is in keeping with what we know of the history and current water management context in the Grand Valley. Table 10 shows that the highly connected nodes in the Grand Valley data set are generally local water utilities (Ute Water Conservancy District, Redlands Water, Grand Valley Water Users Association, Mesa County Water Association, Grand Valley Irrigation Company, Mesa County Irrigation District, Orchard Mesa Irrigation District and Clifton Water District), while city governments (City of Grand Junction, Town of Palisade) are also highly connected.

6.4. Discussion

These analyses together paint pictures of the four study areas that are very much in keeping with expectations based on other sources and an understanding of the areas' historical trajectories. The betweenness statistic applied to the Step 2 results gives an indication about which entities are more central in the discourse about water issues in that area. Flagstaff is a small city in a large region managed by primarily state and federal agencies; the discourse in Flagstaff has a focus on its very large neighbor with pressing water issues (Las Vegas) and on the federal, state and local government agencies that work to shape its water milieu. In Tucson, the central focus of discourse is the city council, with a secondary focus on the institution that is now Tucson's lifeline, the Central Arizona Project, while also integrating discussions from the various outlying towns and a strong emphasis on the city-wide school system. Las Vegas is a large city less secure than Tucson in its water supply and, therefore, pursuing a regional water game; this has implications in the courts (leading, presumably, to the involvement of the Nevada Supreme Court), while other water-specific institutions bridge the gap between these regional connections and the local context. Having begun as a collection of disparate towns, the Grand Valley, despite having a large number of water-specific organization, centers its discourse around the local governments (towns and counties) that must additionally interact.

The graph strength statistic, applied to the reduced data sets of Step 3, shows which of the entities are discussed in conjunction with one another. The story that this gives is somewhat more specific, and it is here that the example from the Grand Valley is the most different. Flagstaff has few entities that must interact at all at the local level; in Tucson, the situation is dominated by the central entity of Tucson Water, and while there are other entities that are all discussed along with Tucson Water, they are far less commonly discussed with each other. This is virtually the same in Las Vegas, but it is quite different in the articles from the Grand Junction Free Press. The central collection of entities in the

Grand Junction data set are shown to be discussed with each other commonly. Our inference is that these entities must interact in their management of water, an inference that is supported by the manual inspection of the original text sources (and other sources, e.g., [30,31]).

Taken together, these analyses provide a convenient entry into the broad picture of water management in the four study areas. The constellation of entities that are revealed from the discussions in the newspaper articles matches the theoretical and empirical studies of water management in which the "ecology of games" is found and through which the dynamics of competition, cooperation and coordination that lead to either friction, adaptive management, or both, are played out. The results of the analyses can be used as either a stepping-off point for a more in-depth analysis using traditional and established techniques or, as we hope, as a basis for a wider simulation modeling exercise. Mapping the structure of the networks of management institutions and actors can be a strong first step in understanding the "ecology of games" and to exploring the advantages or disadvantages of fragmented and varied *vs*. centralized and heterogeneous systems.

7. Conclusions

The results presented here demonstrate that natural language processing and network analyses can capture significant differences between water systems in the four assessed areas. These differences reflect the different management strategies in these areas, from centralized to more fragmented systems. Such results point to the nature and long history of water management in the West of the United States, where this history is evident in communities and their media. Consumers in an area with a fragmented structure may know about their local providers, but not understand how their providers fit into a larger picture; management institutions that have small purviews and roughly peer institutions with which they interact may find that large-scale efforts, including acquiring sources or maintaining infrastructure, are out of their reach or not within their scope. Such institutions may also have more difficulty promoting conservation or other water-use policies, because they cannot unilaterally project a unified message across a wide region; this, in turn, may promote confusion among their consumers about the true nature of their water suppliers and, ultimately, their water supply. In this way the fragmentation of the physical infrastructure, the institutions that control that infrastructure and the collective perceptions of the consumers that interact with that infrastructure may combine to interfere with effective management policies [50].

Fragmented systems may be less effective at adaptive responses when larger changes are evident in the entire hydrologic cycle, where institutional investment in a fragmented system may make changes more difficult [51]. Alternatively, decentralized and fragmented systems may carry advantages: the existence of multiple components in the system can provide multiple paths to legitimate outcomes [15] and divide tasks and responsibilities into manageable components [52]. In many cases, fragmentation is the only plausible and realistic option [52], and the smaller units may be more flexible for addressing local problems [53]. The specific advantages or disadvantages of fragmented *vs*. centralized systems derive from the nature of the challenges being faced. Water management has a number of general characteristics that apply in almost all circumstances, but there may also be specific kinds of challenges (e.g., flooding, drought, contamination) that apply in different ways in different specific situations. The effectiveness of a centralized *vs*. a decentralized structure, or of a particular

kind of network structure, in facing and responding to the specific challenges in a given context or across a collection of contexts that share similar characteristics can be considered an open question, amenable to empirical research or to a modeling approach. The technique identified here could begin to disentangle fragmented water management systems, which may be a necessary step in understanding and assessing how a water management system could begin to be better structured, so that key decision-makers may be better able to implement adaptive responses [54]. The methods presented here contribute to a low-cost, rapid method that can be used (on its own or in conjunction with other approaches) both to inform stakeholders about the nature of their water system and to aid in the construction of more adaptive water management strategies.

The methods presented here offer an example that applies large-scale text mining to understanding water management. The work moves from archived collections of texts to a representation of a real-world phenomenon, connecting public perceptions and discourse, via media, to real-world water system infrastructure and the social institutions that operate it. This provides a new and rapid research technique; this technique offers a useful way to gain new insight into complex water management systems and provides a method of entry into a region's water management system that will find more applications as comparable data sources become more plentiful and more easily accessible. This will assist researchers, analysts and others interested in disentangling the often complex relationships and structures that make up water management. We plan to apply the networks generated here to a modeling study; by creating an agent-based model in which water management agents must interact and negotiate within a centralized or fragmented authority structure, informed by the analyses presented here, we will be able to explore the impacts of these structures on effective management and adaptation. We believe that the kind of analysis presented here will prove particularly useful in an era where water management systems are being reassessed for their adaptive capacity in the face of global-scale change.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. BCS-1114851. This work is supported by the U.S. Department of Energy under contract number DE-AC02-06CH11357. We are grateful to several anonymous reviewers for guidance that added considerable strength to the paper. All errors are, of course, our own.

Author Contributions

The text of this article was written by John T. Murphy, Mark Altaweel, Jonathan Ozik, Richard B. Lammers and Nicholson T. Collier, with contributions by Paula Williams, Drew Cason, Lilian Alessa and Andrew Kliskey; John T. Murphy, Paula Williams, Drew Cason, Mark Altaweel and Andrew Kliskey contributed background research. The development of the new software presented in this article was undertaken by Nicholson T. Collier, Jonathan Ozik, Mark Altaweel and John T. Murphy. Data analysis was performed by John T. Murphy, Mark Altaweel, Jonathan Ozik and Nicholson T. Collier.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Rahaman, M.M.; Varis, O. Integrated water resources management: Evolution, prospects and future challenges. *Sustain. Sci. Pract. Policy* **2005**, *1*, 15–21.
2. Kenney, D.S. *Resource Management at the Watershed level: An Assessment of the Changing Federal Role in the Emerging Era of Community-Based Watershed Management: Report to the Western Water Policy Review Advisory Commission*. Natural Resources Law Center, University of Colorado School of Law: Boulder, CO, USA, 1997.
3. Lubell, M.N.; Robins, G.; Wang, P. Policy Coordination in an Ecology of Water Management Games. Available online: http://opensiuc.lib.siu.edu/pnconfs_2011/22 (accessed on 21 May 2014).
4. Lubell, M. Governing institutional complexity: The ecology of games framework. *Policy Stud. J.* **2013**, *41*, 537–559.
5. McGinnis, M.D. Networks of adjacent action situations in polycentric governance. *Policy Stud. J.* **2011**, *39*, 51–78.
6. McGinnis, M.D.; Ostrom, E. SES Framework: Initial changes and continuing challenges. *Ecol. Soc.* **2014**, *19*. Available online: http://dx.doi.org/10.5751/ES-06387-190230 (accessed on 22 May 2014).
7. Van Meerkerk, I.; Buuren, A.; Edelenbos, J. Water managers' boundary judgments and adaptive water governance. An analysis of the Dutch Haringvliet sluices case. *Water Resour. Manag.* **2013**, *27*, 2179–2194.
8. Hajer, M.; Versteeg, W. Performing governance through networks. *Eur. Polit. Sci.* **2005**, *4*, 340–347.
9. Van Meerkerk, I.; Edelenbos, J.; Klijn, E.-H. Connective management and governance network performance: The mediating role of throughput legitimacy. Findings from survey research on complex water projects in the Netherlands. *Environ. Plan. C Gov. Policy* **2014**, doi:10.1068/c1345.
10. Huitema, D.; Mostert, E.; Egas, W.; Moellenkamp, S.; Pahl-Wostl, C.; Yalcin, R. Adaptive water governance: Assessing the institutional prescriptions of adaptive (Co-) management from a governance perspective and defining a research agenda. *Ecol. Soc.* **2009**, *14*, 1–19.
11. Schneider, M.; Scholz, J.; Lubell, M.; Mindruta, D.; Edwardsen, M. Building consensual institutions: Networks and the national estuary program. *Am. J. Polit. Sci.* **2003**, *47*, 143–158.
12. Bennett, W.L. *News: The Politics of Illusion*; Longman, Inc.: New York, NY, USA, 1983.
13. Alessa, L.; Kliskey, A.; Williams, P. The distancing effect of modernization on the perception of water resources in Arctic communities. *Polar Geogr.* **2007**, *30*, 175–191.
14. Alessa, L.; Kliskey, A.; Williams, P. Forgetting freshwater: Technology, values, and distancing in remote arctic communities. *Soc. Nat. Resour.* **2010**, *23*, 254–268.
15. Korthagen, I.; van Meerkerk, I. The Effects of media and their logic on legitimacy sources within local governance networks: A three-case comparative study. *Local Gov. Stud.* **2014**, 1–24.
16. Gartin, M.; Crona, B.; Wutich, A.; Westerhoff, P. Urban ethnohydrology: Cultural knowledge of water quality and water management in a desert city. *Ecol. Soc.* **2010**, *15*, 36.

17. Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* **2008**, *1*, 1–135.
18. City of Flagstaff Utilities. *Utilities Integrated Master Plan (Draft): Water Resources Chapter- Water History, Demand, Existing Supplies, and Future Water Needs and Recommended Options*; City of Flagstaff-Utilities Division: Flagstaff, AZ, USA, 2011.
19. Logan, M.F. *Fighting Sprawl and City Hall: Resistance to Urban Growth in the Southwest*; University of Arizona Press: Tucson, AZ, USA, 1995.
20. City of Tucson. *Annual Water Quality Reports—2011*; Archive the Official Website for the City of Tucson: Tucson, AZ, USA; Available online: http://cms3.tucsonaz.gov/water/anwqrpts_archive (accessed on 18 December 2013).
21. Southern Nevada Water Authority. *Water Resource Plan '09*; Southern Nevada Water Authority: Las Vegas, NV, USA, 2009.
22. Green, E. Satiating a booming city. Available online: http://www.lasvegassun.com/news/2008/jun/01/satiating-booming-city/ (accessed on 8 November 2013).
23. Green, E. The chosen one. Available online: http://www.lasvegassun.com/news/2008/jun/08/chosen-one/ (accessed on 8 November 2013).
24. Green, E. The Equation: No water, no growth. Available online: http://www.lasvegassun.com/news/2008/jun/15/equation-no-water-no-growth/ (accessed on 8 November 2013).
25. Green, E. Not this water. Available online: http://www.lasvegassun.com/news/2008/jun/22/not-water/ (accessed on 8 November 2013).
26. Green, E. Owens Valley is the model of what to expect. Available online: http://www.lasvegassun.com/news/2008/jun/29/owens-valley-model-what-expect/ (accessed on 8 November 2013).
27. Simonds, W.J. Grand Valley Project, 1994. U.S. Bureau of Reclamation. http://www.usbr.gov/projects//ImageServer?imgName=Doc_1305042485344.pdf (accessed on 21 May 2014).
28. Ute Water Conservancy District; City of Grand Junction; Clifton Water District. *Grand Valley Regional Water Conservation Plan*; Ute Water Conservancy District, City of Grand Junction, and Clifton Water District: Grand Junction, CO, USA, 2013.
29. Shockley, P. Sink or Swim Meeting Wednesday will Shape Palisade's Water Future. *Post Independent*, 30 November 2005. Available online: http://www.postindependent.com/article/20051129/local/111290005 (accessed on 21 May 2014).
30. Water Fight in Palisade. *Daily Sentinel*, 29 November 2005.
31. Mayor: Plan will "Keep Our Water in Palisade." *Daily Sentinel*, 30 November 2005.
32. Analyzing Agents and Aqua. Available online: http://aaa.uchicago.edu (accessed on 21 May 2014).
33. Apache UIMA. Available online: http://uima.apache.org/ (accessed on 18 December 2013).
34. Chisholm, E.; Kolda, T.G. *New Term Weighting Formulas for the Vector Space Method in Information Retrieval*; Computer Science and Mathematics Division, Oak Ridge National Laboratory: Oak Ridge, TN, USA, 1999.
35. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523.
36. Apache OpenNLP. Available online: http://opennlp.apache.org/index.html (accessed on 14 January 2014).

37. Nadeau, D.; Sekine, S. A survey of named entity recognition and classification. *Lingvisticae Investig.* **2007**, *30*, 3–26.
38. Bastian, M.; Heymann, S.; Jacomy, M. Gephi: An open source software for exploring and manipulating networks. In Proceedings of the Third International ICWSM Conference, San Jose, CA, USA, 17–20 May 2009.
39. Gephi, an open source graph visualization and manipulation software. Available online: http://gephi.org/ (accessed on 18 December 2013).
40. Shannon, P. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504.
41. Hu, Y. Efficient, high-quality force-directed graph drawing. *Math. J.* **2005**, *10*, 37–71.
42. Fruchterman, T.M.; Reingold, E.M. Graph drawing by force-directed placement. *Softw. Pract. Exp.* **1991**, *21*, 1129–1164.
43. Wasserman, S.; Faust, K. *Social Network Analysis: Methods and Applications*; Cambridge University Press: Cambridge, NY, USA, 1994.
44. Barrat, A.; Barthelemy, M.; Pastor-Satorras, R.; Vespignani, A. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 3747–3752.
45. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: New York, NY, USA, 2008.
46. Companies we supply services. Available online: http://www.southwesternutility.com/view/193 (accessed on 20 December 2013).
47. Tucson_AMA_Map.pdf Available online: http://www.azwater.gov/AzDWR/WaterManagement/Wells/documents/Tucson_AMA_Map.pdf (accessed on 21 May 2014).
48. Southern Arizona Water Utilities Association—Membership. Available online: http://www.sawua.org/members_logo.html (accessed on 21 December 2013).
49. Water Systems in Pima County, AZ—Toxic Waters—The New York Times. Available online: http://projects.nytimes.com/toxic-waters/contaminants/az/pima (accessed on 21 May 2014).
50. Brown, R.R.; Farrelly, M.A. Delivering sustainable urban water management: A review of the hurdles we face. *Water Sci. Technol.* **2009**, *59*, 839–846.
51. Pahl-Wostl, C. Transitions towards adaptive management of water facing climate and global change. *Water Resour. Manag.* **2007**, *21*, 49–62.
52. Edelenbos, J.; Teisman, G.R. Symposium on water governance. Prologue: Water governance as a government's actions between the reality of fragmentation and the need for integration. *Int. Rev. Adm. Sci.* **2011**, *77*, 5–30.
53. Lubell, M.; Lippert, L. Integrated regional water management: A study of collaboration or water politics-as-usual in California, USA. *Int. Rev. Adm. Sci.* **2011**, *77*, 76–100.
54. Pahl-Wostl, C. The implications of complexity for integrated resources management. *Environ. Model. Softw.* **2007**, *22*, 561–569.