NIH Public Access
Author Manuscript
Int J Comput Models Algorithms Med. Author manuscript; available in PMC 2011 September 30.

Published in final edited form as:
Int J Comput Models Algorithms Med. 2011 ; 2(2): 1–22. doi:10.4018/jcmam.2011040101.

Identifying Temporal Changes and Topics that Promote Growth Within Online Communities: A Prospective Study of Six Online Cancer Forums

Kathleen T. Durant,
Silverlink Communications, USA

Alexa T. McCray, and
Harvard Medical School and Beth Israel Deaconess Medical Center, USA

Charles Safran
Harvard Medical School and Beth Israel Deaconess Medical Center, USA

Abstract

In this paper the authors have extended the methodology for temporal analysis of online forums and applied the methodology to six online cancer forums (melanoma, prostate cancer, testicular cancer, ovarian cancer and breast cancer). The goal was to develop, apply and improve methods that quantify the responsiveness of the interactions in online forums in order to identify the users and topics that promote use and usefulness of these online medical communities. The evolutional stages that gauge when a forum is expanding, contracting, or in a state of equilibrium were considered. The response function was thought to be an approximation of a discussion group's utility to its members. By applying the evolutionary phase algorithm, it was determined that two out of six of the forums are in contracting phases, while four are in their largest growth phase. By analyzing the topics of the influential threads, the authors conclude that cancer treatment discussions as well as stage IV cancer discussions promote growth in the forums. It is observed that the discussion of treatment rather than diagnosis is important to help a cancer forum thrive.

Keywords

Cancer Forums; Influential Members; Methodology; Online Medical Discussion Groups; Social Network Analysis; Temporal Analysis; Topic Identification

1. INTRODUCTION

People join online medical discussion groups to discover medical information, coping strategies as well as support while dealing with their particular disease. Members of a medical discussion group create relationships through the sharing of information. One member poses a discussion topic and other members post relevant text to the topic. This online communication process is known as thread creation. Once created, a thread is a permanent informational resource for all members of an online community.

The utility of a discussion group is its ability to provide the wanted response to a member in a short period of time. Members actively choose to spend valuable time on the medical forum during a very difficult time in their lives (dealing with cancer); this decision will be reversed if the forum does not provide a benefit to its members. Forums that are losing information-seeking members may fail to thrive, since these communities need people to start discussions. Forums not responding to inquiries may also fail to thrive, since people

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

want a response to the questions they pose. People dealing with cancer should be provided online resources that quickly service their informational needs.

Modeling the communicative interactions of an online discussion group must involve a temporal aspect since the active participants as well as the number of interactions between the members varies substantially across time. The majority of the members spend a short period of time contributing to the forum. If other members who behave differently replace members, the experience (experience such as the responsiveness, topics discussed, quality of the information, etc) at the medical forum will vary across time. In order to keep the growth rate as well as the benefit of the online forum at its current state, new members willing to become voluntary online caregivers to other members of the online community must replace non-active members.

This research extends a methodology for assessing the responsiveness, the temporal changes, and the topics that elicit a strong response from an online forum (Durant et al., 2010b). We investigated the composition of different roles that members play within the forum and measure these values at six online cancer forums that vary in size, response rate and topology. We believe this research is relevant to the medical community given the number of health-information seekers that are turning to online resources for their medical informational needs (Fox, 2009).

By providing a methodology as well as metrics for quantifying support at online medical communities, this research provides an initial step for assessing and comparing the quality of support provided by online medical communities. We believe these metrics help develop communication metrics analogous to the communication metrics found in the Consumer Assessment of Healthcare Providers and Systems (CAHPS) program used for assessing the quality of professional medical communication (U. S. Department of Health and Human Services, 2010).

2. RELATED WORK

We have previously defined a methodology (Durant et. al, 2010b) using a phase detection algorithm and response function and applied it to a melanoma forum. We extended the analysis of thread topics in order to compare the topics that are influential in each calendar year as well as for the duration of each cancer forum.

Temporal data models have been analyzed by Leskovec (2007, 2008a, 2008b, 2008c). However, their analysis is not completely applicable to communicative network models. The communicative networks we present in this paper do not follow Leskovec's proposed growth power law. Our communicative networks are similar to Leskovec's models since they do display heavy left tails (but not right tails) for in and out degree distributions. However, our models do not get denser over time as Lescovec's do. Over certain time periods the example networks get sparser. The occurrence of this phenomenon is explained in subsequent sections and is related to the temporary nature of the data elements.

Kostakos (2009) also defined temporal graphs; however his model is concerned with the dispersion of information across time. This is not the focus of this research, since within message boards, information need not pass from member to member to spread throughout the network. Once created, threads are permanently associated with the message board and may be read by any person visiting the message board.

Tang et al. (2009) defined temporal distance metrics between nodes within a temporal graph. Tang et al. (2009) defined a metric to measure the distance between nodes in different time

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

slices. We are primarily interested in understanding the topics and users that are central at each time slice to identify trends across a time frame.

Another relevant study was conducted by Kossinets and Tavel (2006). They analyzed a social network defined by the transmission of email between university members. It is a network similar to a message board forum, where the main difference is the rich collection of actor properties. They measured the changes in network topology over time. They observed that the network properties approach an equilibrium state, an observation that we also observe within our model networks. They are also concerned with measuring the strengths of ties between actors in the network, a concept Durant et al. (2010a) defined as intimacy in a previous paper.

Since the social networks associated with message boards have different characteristics than other studied social networks, most of the interesting research questions associated with them differ from previously posed research questions, making much of the social network research not directly applicable.

3. METHODOLOGY

In this section we describe the data collection process, the temporal scope of objects within an online forum, the growth and response variables, a description of the response function, the social network model, tools and analysis used in this study.

3.1. Forum Data Collection

The website www.cancercompass.com, is an online cancer data source sponsored by the Cancer Treatment Centers of America™. Cancercompass has over thirty cancer forums and dates back to 2001. We harvested the posts, threads and users' data from the melanoma, renal cell, prostate, testicular, ovarian, and breast cancer forums using html parsers. We collected only the publicly available data at the website. The data was collected on May 15, 2010; however, we truncated the new thread collection back to April 30, 2010, giving each forum two weeks to respond to an unanswered thread. The sizes of the collected corpuses are defined in Tables 1 and 2. A description of the data objects follows.

A member is a person who has registered with the discussion forum. A member may have one of several different user types: caregiver, patient, survivor, doctor, nurse, student, researcher and unknown. Members self-assign a user type when they register at the Cancercompass website. A user type typically describes the relationship the user has with cancer. Members with the unknown user type chose not to specify their relationship with cancer. As displayed in Table 1, the number of users within the six communities varies greatly; the breast cancer forum is quite large (3,288 members) compared to the testicular cancer forum (97 members). There are 7,991 members across all six forums.

A thread is created when one member poses a discussion topic and other members post relevant text to the topic. It is a discussion found on a discussion forum; a collection of inter-related posts. A member who poses a discussion to the forum is the creator or the author of a thread. The author of the thread starts a discussion within the online community; a discussion that is open and may be joined by any existing member.

Each column in Table 2 displays the number of threads and posts created by the different user types for a particular cancer forum. Not surprisingly, given the small number of testicular forum members, there are few threads (13) and posts (37) on this forum. The breast forum is the largest forum; however the larger size does not increase the prolificacy of the members. Even though there are 2.37 times more breast forum members than prostate

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

members, they only compose 1.47 times more posts than the total collection of prostate forum members.

3.2. A Forum Objects' Temporal Scope

The classes of objects associated with a forum are: members, threads and the forum itself. Each object within an online forum has a temporal scope. A temporal scope is a time period when the object is defined for the forum. The temporal scope of the forum is defined as the time segment from the point in time when the message board first became available to people on the web (2001) to the point in time when the message board is no longer available on the web (date to deactivate the site) (Durant et al., 2010b).

Threads are *persistent* data objects of a message board. A persistent data object is a data object that, once created, will exist for the remaining temporal scope of the message forum. Members are *ephemeral* data objects of a message board. An ephemeral data object is a data object that has a limited temporal scope. A member exists within the forum while he/she is an active participant of the discussion forum. A member is considered active while he/she is contributing to the forum by creating posts. This definition is a conservative approximation for the existent scope since it terminates the user's scope at the last recorded operation performed by the member. The heuristics allows us to remove inactive members and get a better sense of the size of the active member community.

Objects within a discussion board have an active scope. An active scope is a temporal scope such that the object is producing an action or is being acted upon. A member's active scope is equivalent to its temporal scope. A thread's active scope is shorter than its temporal scope since it is defined as the time period from the creation of its first post to the creation of its last post. Since this analysis occurs prospectively, we can identify the last written post for a thread.

3.3. Social Network Model

A social network consists of actors and ties (Nooy, 2005). The actors are the entities within the network. A tie is a line between two actors; a tie represents a relationship between the two actors. A social network is pictorially represented as a graph where the actors are the nodes and the ties are represented as directional or bidirectional edges. The main goal of social network analysis is detecting and interpreting patterns of social ties among actors (Nooy et al., 2005; Hansen et al., 2011).

Social network graphs attempt to represent the strength of social connections between actors. In our network, we treat the exchanges of posts as an approximation of social ties. The more communication that occurs between two actors, the more intimate the relationship is between the actors. The more intimacy between two nodes, the closer the two vertices will be placed in the graph. The most connected nodes are placed in the center of the graph.

We modeled online forums as a social network with two distinct classes of actors: members and threads. A social network that models different types of objects is called a bimodal graph (Nooy, 2005). A bimodal graph allows us to explore the relationships between the different object classes: threads and members. For example, we can discover clusters or cliques of members who discuss particular topics.

A member node represents a forum member and is created when that user writes his/her first post. A member node represents the member that creates the original communication element or post. A member node may be directly connected to another member node or a thread. A member's data is part of his/her digital footprint since it is created directly by the user (Gantz et al., 2008).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

A thread node represents an online conversation that took place within the message board. A thread node is created when the thread's author writes the first post. Thread nodes are the communication artifacts among the users; they come into existence by users' creating posts. A thread node cannot directly connect to another node; a member node can only connect to it. In other words, all edges of threads are incoming edges. Posts are part of a user's digital footprint since it is data directly generated by the user. However, the data associated with edge creation is part of a user's digital shadow since it is not user-generated data but a description of a member's communication patterns within the forum (Gantz et al., 2008).

The relative thickness of an edge represents the number of directed communication between the two nodes. An edge value greater than one, means a particular member responded multiple times to the same thread or directly communicated multiple times to another member.

Non-directed communication such as questions posed to the community-at-large is modeled as an attribute of the member. It is the inquisitiveness level of the user and is represented by the relative size of the user's node within the network (Durant et al., 2010a). In general, threads provide a mechanism for members to potentially connect, while posts are the manifestation of a connection between two members or between a member and a thread. Within a network, a member node is created because of a member's interest in the forum's topic, whereas an arc is created because of a user's willingness to actively contribute to a thread's discussion. When a member responds to a thread, he/she is making a direct connection to a thread and the thread's author.

3.4. Visualizing the Social Network Model

We visualize the social networks using the Pajek visualization tool (Nooy et al., 2005). We use the Fructerman-Reingold graph-drawing algorithm to represent the social networks (Fructerman et al., 2004). This method models a graph as a mechanical collection of electrically charged rings (the nodes) and connecting springs (the edges). Every two nodes reject each other with a repulsive force, and adjacent nodes (nodes connected by an edge) are pulled together by an attractive force. Over a number of iterations, the forces modeled by the springs are calculated and the nodes are moved within the plane to minimize the total energy of the system (Fructerman et al., 2004).

3.5. Influential Nodes and Topics

In a previous paper, we defined four subclasses of member nodes: members who receive information (consumers), members who provide information (providers), members who receive and provide information (facilitators) and members who do neither (Durant et al., 2010a). We extended the consumer definition to include thread nodes, since a thread node can only be provided information by another member; all thread nodes are consumer nodes.

Since the functionality of these different roles has different influences on the forum, we trace the percentages of these three roles across the time continuum.

We use hub/authority analysis to identify the influential producers, the facilitators and the p-satisfied consumers for the complete network as well as for each year within the time period (Kleinberg, 1999). A hub node is a node that has many outgoing arcs but very few incoming arcs. An authority node is a node that has many incoming arcs but few outgoing arcs (Kleinberg, 1999). Since our network models data transfer, the hubs are the users providing the most information, the most active producers. Authorities are the nodes (either threads or members) that receive a large response from the network, the p-satisfied consumers (Kleinberg, 1999). Users identified as both a hub and an authority are nodes that have many incoming and outgoing arcs, the most successful facilitators in communication. Facilitators

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

pass information and are passed information; they encourage communication by communicating in a conversational manner (Durant et al., 2010a).

We measure the frequency of bigrams and unigrams within the topics of the p-satisfied threads, since this can identify the reoccurrence of a topic being discussed heavily within the forum. The most frequently occurring unigrams and bigrams within the p-satisfied threads are the topics that elicit the strongest response from the forums and hence are the topics that encourage growth within the forum. We also analyze the unigrams and bigrams across the six forums in order to identify common themes that occur within each forum. These themes may be represented by varying unigrams and bigrams.

3.6. Growth, Response, and Activity Duration Metrics

We measured the activity duration for members and threads in order to understand the temporal aspect of objects within the forum. Forums where the members have longer membership durations have more opportunity to form relationships with other members. Shorter membership durations mean the feel of the forum is more likely to be different across time since the communicators are changing more rapidly than a forum with longer membership duration. The length of the thread's activity duration also affects the feel of a forum. Longer thread durations mean the threads' topics appeal to members for a longer period of time. Shorter thread durations mean the threads' topics have less appeal as they age. The object duration time represents the turnover rate of the forum.

We measured the growth metrics for a forum: in users, posts, threads and connections. These metrics are collected for each month, each calendar year as well as for the complete timeframe. All four of these growth measurements affect the feel of a network. Two networks that contain the same number of members but one network has more connections between the members will have a more intimate feel than a forum with fewer connections. We also compared the interconnectedness of the nodes given the number of posts. This analysis is performed per calendar year.

3.7. Response to a Request Function

We extended the response function (Durant et al., 2010b) to take into account the number of requests, the number of responses actually given to a thread as well as the work requested during a specific time period. By adding these extensions, we created a metric that can be compared across the forums.

Within the response function we wish to reward a forum for threads that are answered quickly, penalize heavily for threads not answered at all and penalize moderately for threads answered with a delay. We also want to reward forums that provide multiple responses to a thread.

We define the work request function WR to be the number of created threads during a particular time period i. The Work performed function WP is defined for each time period i, where i ranges over the set of Natural numbers N. The work performed function WP for a particular time period i is defined as the following:

$$WPi=(Ai*B) - (Wi) - (Ui*P)+(Ri*B/F)$$

Ai is the number of threads written in time period i that receives an initial response during the current or a future time period. B is a bonus factor awarded to the threads that receive a response. Wi is the wait duration for each of the threads (measured in days) for the time period i. Ui is the number of threads written in time period i that has not received a response.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

P is defined as the penalizing factor. Threads that do not receive a response are penalized by the P factor. The sum of Ui and Ai is the total number of threads that were authored in time period i. Ri is the total number of response posts written in time period i. We multiplied it by a fraction of the bonus factor since we believed the first response is more important than the latter responses.

For our experiments, we set both the bonus factor B and the penalizing factor P to a multiple of the maximum response wait time for the time range (2,644 days).

$$P=B=\text{argmax}\,(Wi)*Fi \in N$$

In our equation, i ranges from 1 to n where n is the number of time windows for the time frame. We set the Factor F to 3. This means the penalizing factor for not receiving a response to a thread, is three times the longest wait on the six forums.

We defined the Response to Work Request Function for a time period i to be defined as:

$$R=WPi/WRi \text{ for all time periods } i \in N.$$

3.8. Stage Identification

We wished to identify the different evolutional stages of the six online forums. We defined an evolutional stage as a time period where the growth variables are statistically indistinguishable (Durant et al., 2010b). We segmented the eight year time period using 6 months time windows. We performed Kruskal-Wallis analysis on the growth variables using Bonferroni correction to determine the time periods where the growth variables vary from the other time periods.

4. RESULTS

By calculating the growth variables and applying the stage detection algorithm to these values, we identified the evolutionary stages for each of the six forums. The evolutionary stages allowed us to compare the growth progressions for the six forums. This reveals the evolutionary stages and the growth peaks. By calculating the response to request function for each of the forums, we identified the response associated with each of these phases. Since the response function is a value than can be compared across the six forums, we were able to determine which forums were servicing their requests in a more timely and profuse manner.

By applying hub/authority analysis to the different time periods, we determine p-satisfied consumer threads that represent the topics that elicit strong responses from each of the forums.

4.1. Growth Metrics

Figure 1 displays the thread growth for each 6-month time period on the six discussion board from July 2001 to April 2010. In Figure 1 we see slow thread growth measures for the first few years of each of the forums (Jul 1 through Jul 04). From Jan 06 to Apr 10 we see steadier growth in all but the testicular forum. We do see a decline in growth after January 2009. The last time period contains data for only 4 months of data rather than 6 months.

We present the growth in connections for the six forums. In Figure 2, the x-axis represents the calendar years from 2001 to 2010. The y-axis represents the average number of node

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

connections within each forum for the particular time period. The size of the bubble represents the number of active nodes during that time period. Higher bubbles are networks that are more interconnected (average number of neighbors to a node). Larger bubbles represent networks that have relatively more active members. In 2005, all forum members have approximately the same number of connections even though the sizes of the forums vary. In 2007, we see the prostate and the renal–cell cancer forums continue to increase the number of connections between users. The testicular forum continues to decline in number of average connections as the time period increases, for most of the time period the average number of connections is 0. This fact is represented by the testicular forum not having a bubble associated with a time period (such as 2002–2005 and 2010).

4.2. Response Function

We measured the overall work response function to all work requests for each of the six forums. The work response function is a linear combination of the variables that represent the positive and negative experiences within the forum.

Figure 3 displays the results of the overall response function for each of the 16 6-month time periods for the six forums as well as the median response. We see in the first years of the forum (Jul-01 through Dec-04) low responsive values for all of the forums but the ovarian cancer forum. The median response peaks in the time periods from January 2008 through December 2008. The time period from Jan 2007 to Jun 2009 sees the greatest growth in the number of threads across the forums. For this time period, all forums except for the testicular forum are similar in their responses to a request. This finding shows that the forums could continue to handle the load requested even during times of growth.

In Figure 3 we see the testicular forum has fewer time periods when it is performing above 0 than below or equal to 0. The experience at this forum is not positive for a member seeking advice.

The median line in Figure 3 allows us to compare the work performed at the different forums during the different time periods. The renal cell forum was not performing at the level of the other forums until Jul 2005. Once past that date, the forum performs above the median. The ovarian forum was relatively more productive than the other forums before Jan 2005; it performs below the median after Jan 2005 until July 2008. The breast cancer and the prostate cancer forums consistently perform close to the median in all time periods. The melanoma forum performs inconsistently before Jan 2005, after that date the performance is more consistent.

4.3. Activity Duration Metrics

We measured the activity duration metrics of the six forums using the complete timeframe as the timeframe window. We compared the temporal scope of member and thread objects to determine if members' activity periods last longer than the typical activity periods of a thread. Our results show that a thread's activity period is, in general, longer than a member's activity period, demonstrating the persistent nature of a thread. Members who are not active within the same time period can contribute to a thread, improving the informational resources of the forum. A Shapiro-Wilk test reveals that each activity duration variable is not uniformly distributed (p-value < 0.0001). Given this finding we present the median and the 95% confidence interval for each of the median measures.

Table 3 shows that on all six forums more than half of the members are active for only one day, displaying the dynamic environment of these forums. However, if we examine the duration periods for the top quartile for each forum, we see quite a variation in this activity period for these long-term members. The testicular forum has 75% of its members being

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

active for only one day; it is the only forum whose thread duration is comparable to the member's duration length. All other forums have at least a 6 fold increase in thread length duration compared to user length duration, displaying the permanent nature of the thread object.

The testicular forum is the only forum whose thread duration is comparable to the forum's user duration length. In general, the testicular forum members are not extending the permanent informational resources of the forum leading to low thread activity levels.

4.4. Stage Identification

We applied the stage identification algorithm to the six forums and identify four similar and two different evolutionary phase progressions. The stage identification algorithm determines significant statistical changes in the number of posts from one time period to the next (Durant et al., 2010b).

All six forums have only three distinct phases defined. These distinct phases are: a smaller sized phase (phase 1), a transitional sized phase (phase 2) and a larger sized phase (phase 3). For the six forums, the smaller sized phase ranges from two and a half to four years. The transition phase ranges from one year to two years and the maximum growth phase ranges from six months to four years. The forums also differ from each other in the progressions they make through these stages as demonstrated in Figures 4 through 9.

Stage identification gives us a different view of the forum than the Response function. The definition of each identified stage is specific to a particular forum. Even though we are comparing the evolutionary stages among the forums the actual amount of response and work performed within each of the three distinct phases may vary among the forums.

In Figures 4 through 9, stage 0 represents the time period when the website has gone live but we have seen no activity on this particular forum. Stage 1 is the slower growth period, stage 2 is the transition period and stage 3 is the larger growth period. Figures 4 through 9 plot the stage value for each 6-month time period starting July 2001 and ending Apr 2010. We see the forums went live in different time periods. The breast cancer forum is the first forum to see activity.

We see similar step progressions through the stages for the melanoma, renal cell, prostate, and ovarian cancer forums. All four forums are currently in phase 3; the largest growth phase.

The testicular forum spends most of its time periods in the slow growth stage (stage 1). It circles back to stage 1 during two different time periods (January 2008 and July 2009). It spends only 1 6-month time period in the larger growing time stage (July 2007). It is the only forum that chronologically transitioned from its slower growth stage to its faster growth stage. It does not enter stage 2 until after it has fallen from stage 3 to stage 1. It only remains within the transition phase for a year (July 2008–June 2009); then it falls back to the slow growing stage.

The breast cancer forum is also not currently in stage 3; it has fallen back to the transition growth period. Within stage 3, on average the breast cancer forum created over 200 posts a month; the forum could not sustain this fast paced growth rate beyond July 2009.

4.5. Social Network Model

For each forum, we visualize the most active calendar year for growth. Within Figures 10 through 15, we use both shape and color to distinguish the different user types: red square =

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

patient, blue circle = caregiver, salmon square = survivor, green circle = doctor/nurse, and yellow triangle = undefined user type.

We limit the model to active members for the timeframe. Figure 10 shows the connections for the 310 active users for 2008 for the melanoma forum. The network spreads to use all available space.

The renal cell cancer forum's most active calendar year is 2009. It has 329 active members. The number of members within Figure 11 is comparable to the melanoma forum pictured in Figure 10; however, the nodes are more interconnected as demonstrated by the increase number of edges between the nodes.

The prostate forum is most active during 2008. It has 486 active members. Its high level of interconnectedness is visualized by the denseness and the number of edges within the graph.

The testicular cancer forum is most active in the year 2007; it has 35 active members. Figure 13 demonstrates that the forum does not have a focal point but consists of small clusters of members that communicate only among each other.

The ovarian cancer forum is most active in the year 2008; it has 439 active members. Its size in members is similar to the prostate cancer forum however its topology is quite different since it does not have the high level of interconnectivity as the prostate forum. The ovarian cancer network's topology is similar to the renal cell cancer forum however; it has a few edges with extremely high values. This is represented by the large arrows within the graph.

The breast cancer forum is most active within the year 2008; it has 999 active members. There are many nodes that are not connected to another member. These singleton nodes represent a member that have posed a question to the breast cancer forum and has not received a response.

4.6. Hub Authority Analysis

We applied hub authority analysis to the different forums to identify the most influential nodes (threads and users) for the different calendar years. We limit influential nodes to the top 5% of the active network for each calendar year. If the number of active nodes for a calendar year is fewer than forty, we merge the temporal graph with the next year's temporal graph and perform hub authority analysis on the combined graph for the two years. We use the Pajek toolkit's (Nooy et al., 2005) implementation of hub authority analysis for this study.

We also performed hub authority analysis on the total time period since the influence of some nodes may cross years. The result difference in these two temporal views allowed us to approximate the importance of a particular time period for the overall growth of the forum.

We measured the frequency of bigrams and unigrams within the topics of the p-satisfied threads, since this can identify the reoccurrence of a topic being discussed heavily within the forum. The most frequently occurring unigrams and bigrams within the p-satisfied threads are the topics that elicit the strongest response from the forums and hence are the topics that encourage growth within the forum.

4.6.1. Melanoma Forum—We analyzed the results of the identified p-satisfied consumer nodes and influential providers identified via the hub authority analysis. We determined that p-satisfied consumer threads, containing the bigrams 'stage IV' and 'metastatic melanoma'

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

elicited a stronger than average response from the health information providers of the forum. Treatment methods such as interferon (also known as interferon alpha2b) and Il-2 (also known as Interleukin 2, Proleukin®, aldesleukin) are top unigrams for the p-satisfied consumer threads.

Once facilitator nodes appear within the social network, they continue to be an identified influential class of nodes for the following time periods within the third evolutional stage.

Interestingly, the actual members identified as facilitators vary from time period to time period. There were 12 facilitators identified for the total time period and 10 identified for the years 2006–2010. Out of the 10 facilitator nodes identified, none appear in more than one year. Eight out of 12 facilitator nodes for the total period are not found within the yearly time periods.

4.6.2. Prostate Forum—The top most occurring bigram for the prostate forum is: ‘PSA level’ and the top occurring unigram is ‘HIFU’, high-intensity focused ultrasound. Other traditional treatments are also represented, as well such as Casodex® (also known as bicalutamide), Lupron® (also known as leuprolide), (Chu & DeVita, 2010) imrt (intensity modulated radiation therapy) (Ko et al., 2008) and orchiectomy. Some alternative treatments are mentioned such as prostasol and green tea.

Hub authority analysis identified facilitators for most years. The prostate cancer forum has influential nodes engaging in conversational communication every year, except for 2004.

4.6.3. Renal Cell Forum—The p-satisfied consumer threads for the renal cell cancer forum are discussing Sutent® (also known as sunitinib) (Chu & DeVita, 2010) and its side effects. It is the top most occurring unigram in the p-satisfied consumer threads. Other therapy methods are also mentioned such as sorafenib (also known as Nexavar®), Axitinib® (also known as AG013736), Torisel® (temsirolimus), Interleukin 2, as well as particular clinical trials. Nephrectomy is mentioned less frequently and metastatic cancer and stage IV is directly mentioned infrequently.

4.6.4. Testicular Forum—The p-satisfied consumer threads for the testicular forum are focused on diagnosis and advice. Metastastic cancer is mentioned in one out of four identified p-satisfied consumer threads. Given the size of the forum, there are very few bigrams and unigrams that occur within the topic titles more than once.

There are 4 influential provider nodes. None of them span time segments. Facilitators do not occur for any of the time segments; however 3 were identified for the total timeframe. One of the 3 was identified as an influential provider for a particular timeframe.

4.6.5. Ovarian Cancer—The p-satisfied consumer threads for the ovarian forum are discussing chemotherapy and the ‘cancer agent 125’ test used to evaluate ovarian cancer treatment. The p-satisfied threads are also discussing methods to administer chemotherapy such as intraperitoneal chemotherapy; a method reserved for stage 3 and stage 4 cancer. For the most part, the threads are not using specific drug therapy names, the rarely mentioned treatment agents are: Doxil® (doxorubicin) and Gemzar® (gemcitabine) (Chu & DeVita, 2010). This forum is discussing specific forms of ovarian cancer such as: germ cell tumor and granulosa cell tumor.

Some p-satisfied consumer threads for the ovarian cancer forum are less informational and more relationship-based; they are focused on the clinical progress of one particular forum

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

member (the author of the thread). This type of p-satisfied consumer thread is not identified as influential within any of the other forums.

Facilitator nodes appear in 2005 and continue to be an influential node for the ovarian cancer forum for each following time period. This time period spans the transitional and fast growing phase.

4.6.6. Breast Cancer—The p-satisfied consumer threads for the breast cancer forum are discussing traditional and alternative treatments, stage IV, metastases, radiation and mastectomy. The most frequently occurring unigram within the topics of the p-satisfied consumer threads is ‘Arimidex®’ (anastrozole); ‘Aromasin®’ (exemestane) is also within the top occurring unigrams. Other treatments discussed are: tamoxifen, Femara® (letrozole) and the generic term chemotherapy (Chu & DeVita, 2010). The bigram ‘side effects’ is one of the top occurring bigrams as well as ‘alternative treatment’. Various forms of the word metastases also appear as a top unigram.

Like the prostate cancer forum, the breast cancer forum has facilitators appearing in each evolutionary phase. Only years, 2001 and 2002, do not have facilitator nodes identified.

5. DISCUSSION

The goal of this research is to refine and apply a methodology to model the growth and responsiveness of an online discussion board. This methodology identifies the evolutional stages of an online forum. The evolutional stages of a forum shows when a forum, is growing, contracting or in a state of equilibrium. Knowing the evolutionary stage of a forum could be useful to online information-seekers when they are choosing which online forum to join. It is also a useful tool for forum site managers monitoring the utility of an online forum.

This methodology determines the evolutionary stages by defining and measuring response metrics for different time windows. We use these metrics to compare the responsiveness at different online forums. The metrics provide insight into the utility of an online forum.

By applying this methodology to six cancer forums we identified an evolutional phase progression similar in the melanoma, renal cell and prostate cancer forums. The stages for these forums continue to grow as time proceeds, hitting points of equilibrium that define the three phases. The breast cancer forum also follows this same phase progression until July 2009. At this point the growth falls back to the level associated with the transition phase. This fallback may be acceptable for the forum, given the actual size of the forum or may signal that an intervention that stimulates growth may be needed for the forum. The testicular forum spends most of its time in the slow growth phase; it spends only 6 months in its fast growing phase. Both users and threads are on average only active for one day, the forum would benefit from an intervention to promote growth.

Our methodology identifies the topics that elicit a strong response from a forum. This knowledge can be used to determine a forum’s expertise or the topics that are of interest to the forum’s community. This knowledge can then be used by online information-seekers when trying to determine which online forum to pose their question to. These metrics also identify the online forums that are answering questions in a timely fashion.

Hub authority analysis identified discussions focused on treatment as an influential topic on the 5 thriving cancer forums. The representation of the treatment concept varied from forum to forum. Prostate cancer forum members were discussing high intensive frequency ultrasound; melanoma cancer forum members were discussing interferon and interleukin-2,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

breast cancer patients were discussing Arimidex®, ovarian patients were discussing chemotherapy, and renal cell cancer members were discussing Sutent®. Only the slow growing testicular forum did not include a treatment concept within the topics of its influential thread's topics.

Another important topic that is represented in the influential threads' topics is discussions relevant to dealing with advanced stages of a particular cancer. We see this topic within all of the cancer forums but the testicular cancer forum. Within the melanoma forum, members seeking information on 'stage IV' and 'metastatic melanoma' account for 43% of the p-satisfied thread nodes. The ovarian cancer forum responds to discussions on 'intraperitoneal chemotherapy' a technique used for treating stage 3 and stage 4 ovarian cancer. The prostate cancer forum discusses Provenge® (sipuleucel-T) heavily before the FDA's approval of the medication, a treatment for advanced metastatic prostate cancer. The number one word found within the influential threads of the renal cell cancer forum is Sutent®; a treatment for advanced renal cell cancer. The breast cancer forum's most frequently occurring word is Arimidex®; an aromatase inhibitor used for treating advanced breast cancer in post menopausal women. We believe members seeking health information on these topics are patients or caregivers for the sickest cancer patients associated with the forum. These people need timely support from the forum. With these treatment discussions falling into the most influential threads, information-providing members are more likely to reach out to the sickest members of the online community.

Tests that track the reoccurrence or progression of a cancer during treatment are also highly ranked within the influential threads of the ovarian and prostate cancer forums. Representations of the screening, cancer agent 125, used for tracking progression or reoccurrence of ovarian cancer appears within 17% of the influential threads' topics. Within the prostate forum's influential threads, PSA (prostate specific antigen test) is the topmost occurring unigram, it occurs within 27% of the most influential threads.

There are other insights revealed about individual forums while analyzing the most influential threads. The prostate and the breast cancer forum are the only forums that have alternative treatment concepts among the influential threads. Another interesting finding is specific to the ovarian cancer forum. Some members use the forum as a blog to post their day to day dealings with cancer. Having this type of thread appear among the influential threads means the ovarian cancer forum is a community that shares experiences as well as knowledge with each other. This thread class demonstrates members emotionally supporting one another.

Within the renal cell cancer forum, we see information-seeking members actually posing questions directly to a member that has been identified as an influential facilitator by our methodology. Identifying this type of thread, means that members are aware of the informational providers within their community and are actively seeking advice from these members. Having this type of thread appear within the list of influential threads means that the information-seeking member received an above average response from the forum. Their technique of addressing a question directly to an influential provider provided an above average response.

There are other insights revealed about individual forums while analyzing the most influential threads with respect to the response to request function The ovarian cancer forum performs extremely well when the forum is very small both in size of users and requests. (Jan–Dec 2003). If we look at the bubble representing the ovarian forum in Figure 2, we see, on average, the members have more connections than the other forums for this timeframe. However, it loses this advantage during the next few calendar years. As the forum grows, it

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

cannot maintain the high number of connections. It hits another local maxima in January 2009. We look at the threads associated with both these time periods and identify threads that have a relationship and experiential feel to them. The threads are authored by members who are reaching out to an already existing group of friends as well as any other member willing to chat on the forum. The peak in January 2009 is due to a thread labeled, the teal warriors, a Facebook® group of women linked together because of their experiences with ovarian cancer. The teal warriors from Facebook® all decided to join Cancercompass to extend their community.

6. CONCLUSION

We have presented a methodology for assessing response and identifying the evolutional stages of an online discussion board. This methodology allows online forum owners to understand and assess the communication capacity of the forum. This information will help forum owners and web site administrators to assess and adjust the level of online support for the forum. This information is also valuable to forum members since many members might boost their activity level to increase the benefit the forum provides to the forum community. Lastly, this information is valuable to online health seekers since this information would help them decide which online medical forum to join.

By examining the influential threads within the six cancer forums, we identified discussions focused on treatment as a necessity for a thriving online cancer forum community. Aspects of treatment involved, general treatment, treatment specific to stage 3 and stage 4 cancer, screenings to monitor cancer progression/reoccurrence, alternative treatments, and treatment procedures are the influential treatment subtopics found within the forums.

Acknowledgments

This work was supported by the National Library of Medicine Grant T15-RFA-LM-06-001. Any opinions, findings, and conclusions or recommendations expressed within this material are the authors' and do not necessarily reflect those of the sponsor. We would like to thank the Cancer Treatment Center of America® for the use of their publicly available data.

REFERENCES

Chu, E.; DeVita, VT, Jr. Physicians' cancer chemotherapy drug manual. Boston, MA: Jones and Bartlett Publishers; 2010.

Durant KT, McCray AM, Safran C. Social network analysis of an online melanoma discussion group. Proceedings of the AMIA Clinical Research Informatics Summit. 2010a March.:6–10.

Durant, KT.; McCray, AM.; Safran, C. Modeling the temporal evolution of an online cancer forum; Proceedings of the First ACM International Conference on Health Informatics; 2010a November. p. 356-365.

Fox, S.; Jones, S. The social life of health information. Washington, DC: Pew Internet and American Life Project; 2009.

Fructerman TM, Reingold EM. Graph drawing by force-directed placement. Software, Practice & Experience. 1991; 21(11):1129–1164.

Gantz, JF.; Chute, C.; Manfrediz, A.; Minton, S. The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011. 2008. p. 6-9.Retrieved from http://www.scribd.com/doc/2309354/The-Diverse-and-Exploding-Digital-Universe-Executive-Summary

Goodwin PJ, Leszcz M, Ennis M, Koopmans J, Vincent L, Guther H. The effect of group psychosocial support on survival in metastatic breast cancer. The New England Journal of Medicine. 2001; 345:1719–1726. [PubMed: 11742045]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Guo, L.; Tan, E.; Chen, S.; Zhang, X.; Zhao, Y. Analyzing patterns of user content generation in online social networks; Proceedings of the 15th ACM SIGKDD International conference on knowledge discovery and data mining; 2009. p. 369-378.

Hansen, DL.; Shneiderman, B.; Smith, MA. Analyzing social media networks with NodeXL Insights from a connected world. San Francisco, CA: Morgan Kaufman; 2011.

Kleinberg JM. Hubs, authorities and communities. ACM Computing Surveys. 1999; 31(4)

Ko, AH.; Dollinger, M.; Rosenblaum, EH. Everyone's guide to cancer therapy. Kansas City, MO: Andrews McMeel; 2008.

Kossinets G, Tavel P. Empirical analysis of an evolving network. Science. 2006; 311(88):88–90. [PubMed: 16400149]

Kostakos V. Temporal graphs. Physica A: Statistical Mechanics and its Applications. 2009; 388(6): 1007–1023.

Leskovec, J.; Backstrom, L.; Kumar, R.; Tomkins, A. Microscopic evolution of social networks; Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2008a. p. 462-470.

Leskovec, J.; Horvitz, E. Planetary-scale views on a large instant-messaging network; Proceedings of the 17th International Conference on World Wide Web; 2008b. p. 915-924.

Leskovec, J.; Lang, K.; Dasgupta, A.; Mahoney, M. Statistical properties of community structure in large social and information networks; Proceedings of the International Conference on World Wide Web; 2008c. p. 695-704.

Leskovec, J.; McGlohon, M.; Faloutsos, C.; Glance, N.; Hurst, M. Cascading behavior in large blog graphs; Proceedings of the SIAM Conference on Data Mining; 2007.

Lieberman M, Golant M, Giese-Davis J, Winzlenberg A, Benjamin H, Humphreys K. Electronic support groups for breast carcinoma: A clinical trial of effectiveness. Cancer. 2003; 97(4):920–925. [PubMed: 12569591]

Nooy, W.; Mrvar, A.; Batagelj, V. Exploratory social network analysis with Pajek. Cambridge, UK: Cambridge University Press; 2005.

Norton TR, Manne SL, Rubin S, Carlson J, Hernandez E, Edelson MI. Prevalence and predictors of psychological distress among women with ovarian cancer. Journal of Clinical Oncology. 2004; 22(5):919–926. [PubMed: 14990648]

Spector, AZ. Achieving application requirements. In: Mullender, S., editor. Distributed systems. New York, NY: ACM Press; 1989. p. 19-33.

Tang, J.; Musolesi, M.; Mascolo, C.; Latora, V. Temporal distance metrics for social network analysis; Proceedings of the 2nd ACM Workshop on Online Social Networks; 2009. p. 31-36.

U. S. Department of Health and Human Services. Consumer assessment of healthcare providers and systems. 2010. Retrieved from www.cahps.ahrq.gov

Winzelberg AJ, Classen C, Alpers GW, Roberts H, Koopman C, Adams RE. Evaluation of an internet support group for women with breast cancer. Cancer. 2003; 98(5):1164–1173. [PubMed: 12599221]

Biographies

Kathleen T. Durant is an Analytics Manager at Silverlink Communications, a healthcare communications company that connects with individuals in personalized and relevant ways to help them make better health decisions. Prior to this appointment she was a National Library of Medicine Research Fellow at Harvard Medical School within the Division of Clinical Informatics at the Beth Israel Deaconess Medical Center. Her chosen research domains are: topic and sentiment analysis, social network analysis, health communication, and cancer. She received her PhD in Computer Science from the School of Engineering and Applied Sciences at Harvard University. She is an advisor to the MD Idea Lab (MDiLab), an informatics lab that builds mobile and Internet healthcare prototypes and applications. She has previously worked at Thinking Machines Inc and GenRad Inc.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Alexa T. McCray is Associate Professor of Medicine at Harvard Medical School and the Department of Medicine, Beth Israel Deaconess Medical Center. She is a co-Director of the Center for Biomedical Informatics at Harvard Medical School as well as the Principal Investigator of the NIH-funded Boston-Area Biomedical Informatics Research Training Program. She was recently asked to serve on a National Academy of Sciences standing Board on Research Data and Information, and she was elected this past year as President-Elect of the American College of Medical Informatics. Her chosen research domains are: biomedical informatics, including autism spectrum disorders phenotype-genotype correlations, biomedical ontologies, and health communication. She is the former Director of the Lister Hill National Center for Biomedical Communications, a research division of the National Library of Medicine (NLM) at the National Institutes of Health. She received her Ph.D. in Linguistics from Georgetown University.

Charles Safran is a primary care internist who has devoted his career to improving patient care through the use of informatics. He is Chief of the Division of Clinical Informatics, Beth Israel Deaconess Medical Center and Associate Professor of Medicine Harvard Medical School. Dr. Safran is co-Editor of the *International Journal of Medical Informatics* and on the Health on the Net (HON) Foundation Council. He has helped develop and deploy large institutional integrated clinical computing systems, electronic health records, clinical decision support systems to help clinicians treat patients with HIV/AIDS and most recently personal care support systems for parents with premature infants. He has over 150 publications and speaks to national and international audiences and has testified for the U.S. Congress on Health IT. He graduated cum laude in Mathematics and hold a Masters degree in mathematical logic and a Doctor of Medicine all from Tufts University.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

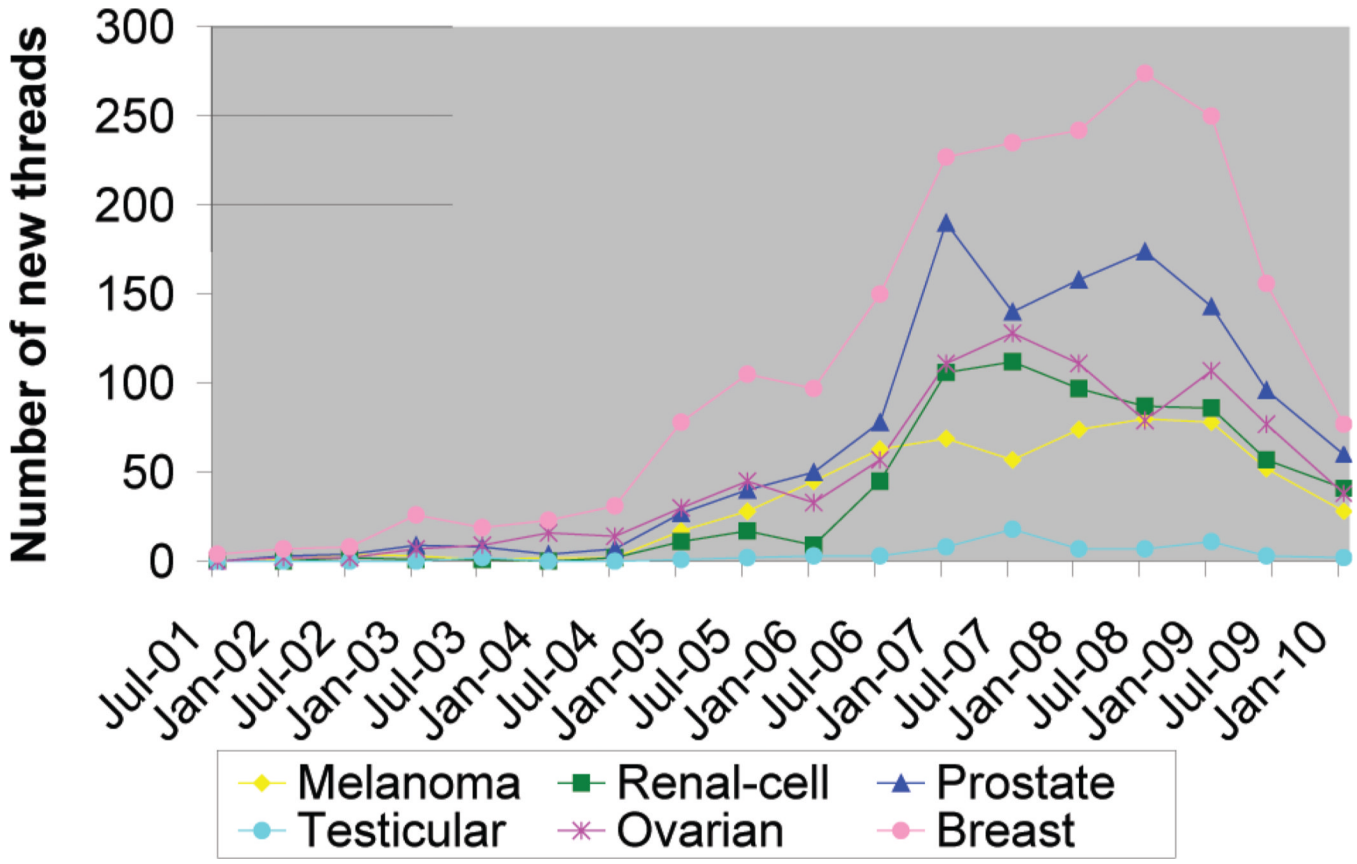

Figure 1.
Growth of new threads for each 6-month time period for the six forums

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

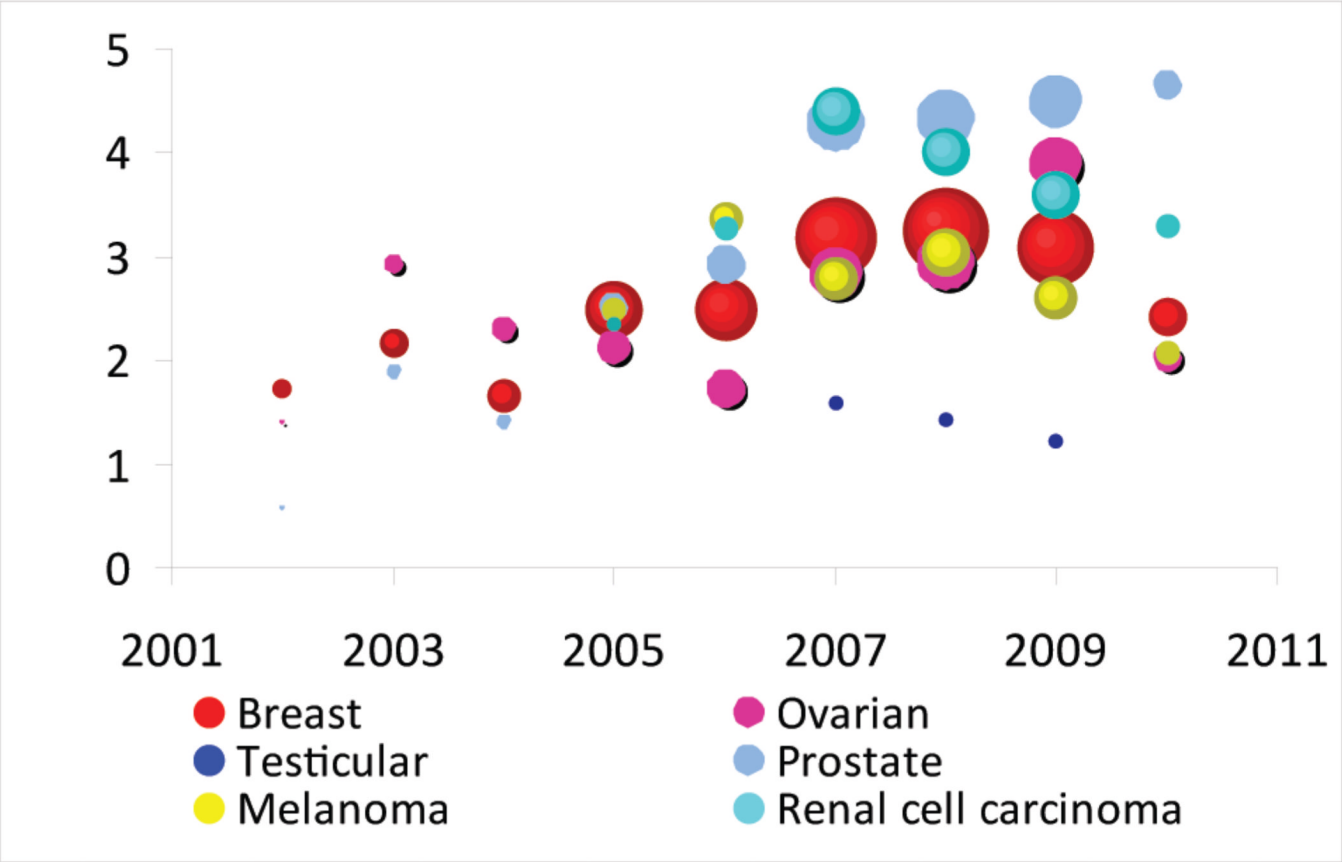

Figure 2.
The growth in active users and user connections from 2001 to 2010

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

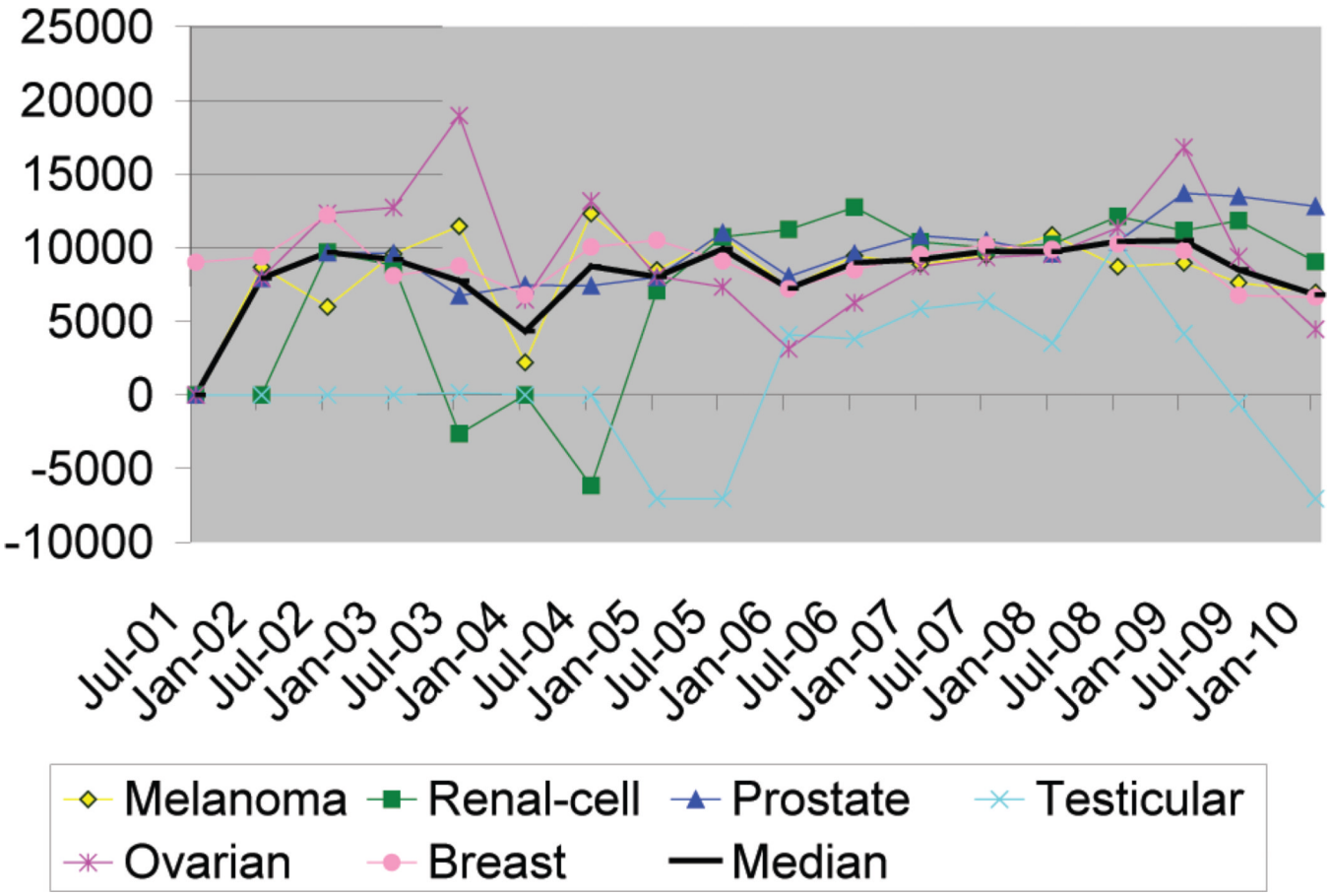

Figure 3.
Response function per work request for each 6-month time

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

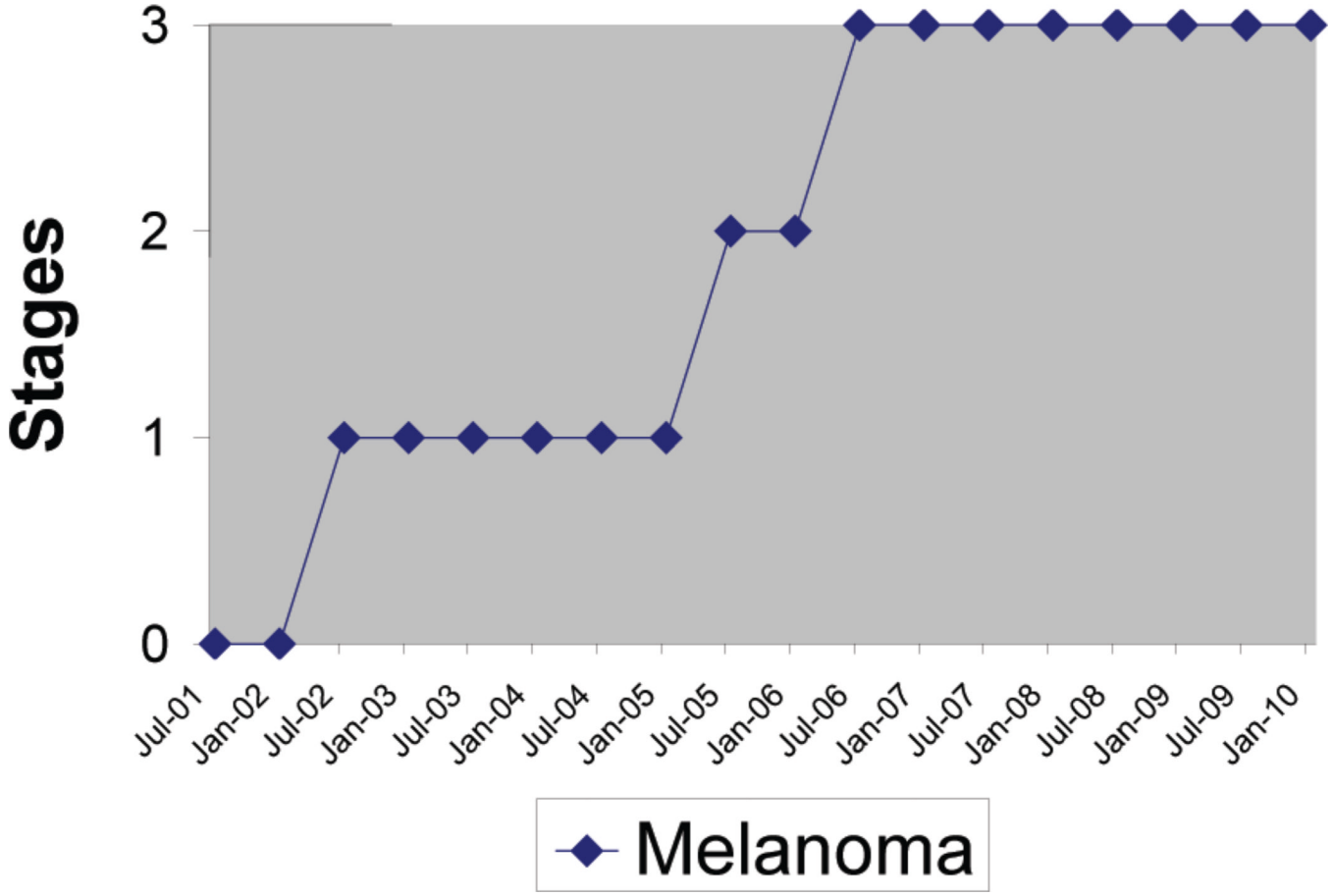

Figure 4.
Melanoma forum stage progression

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

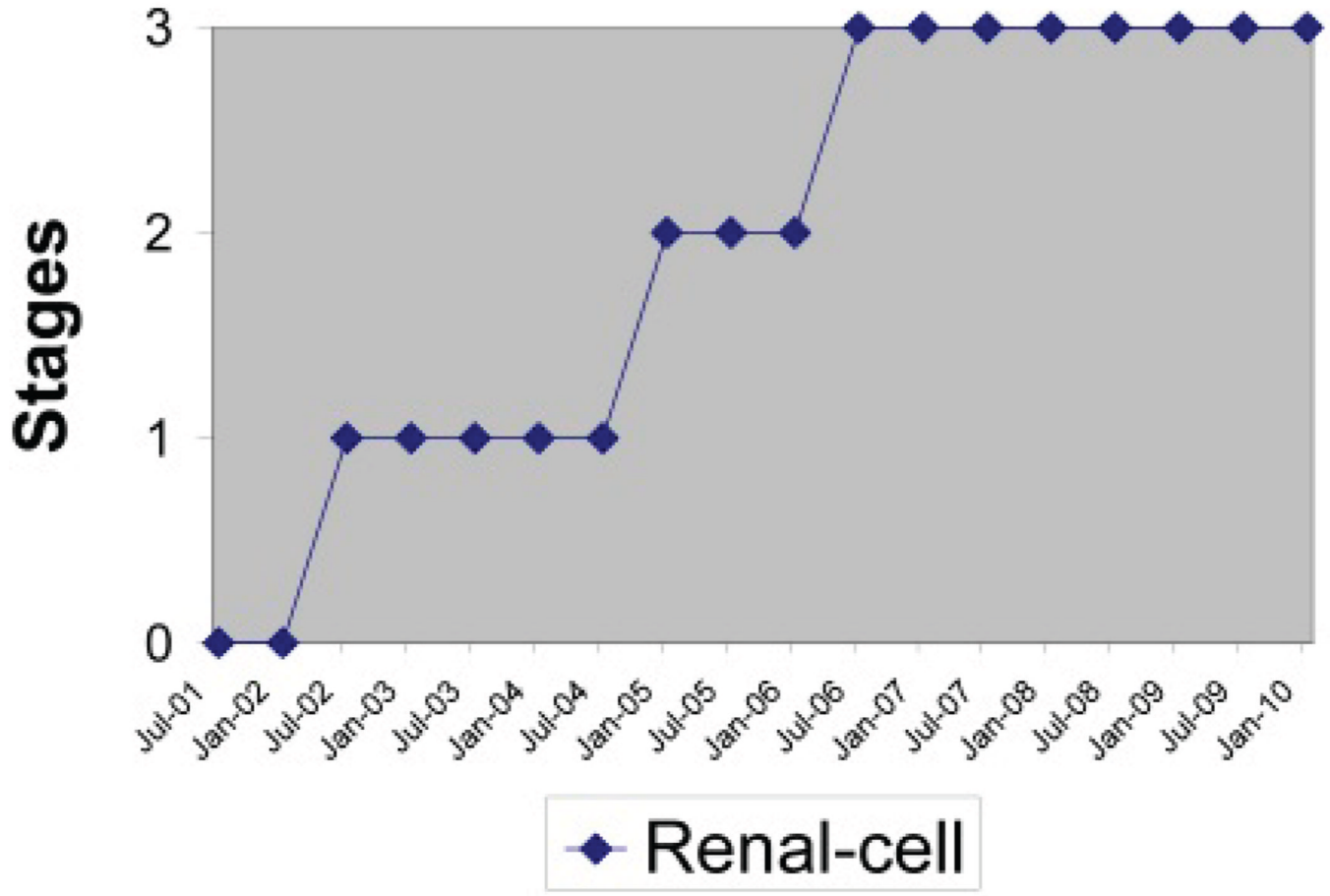

Figure 5.
Renal cell cancer forum stage progression

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

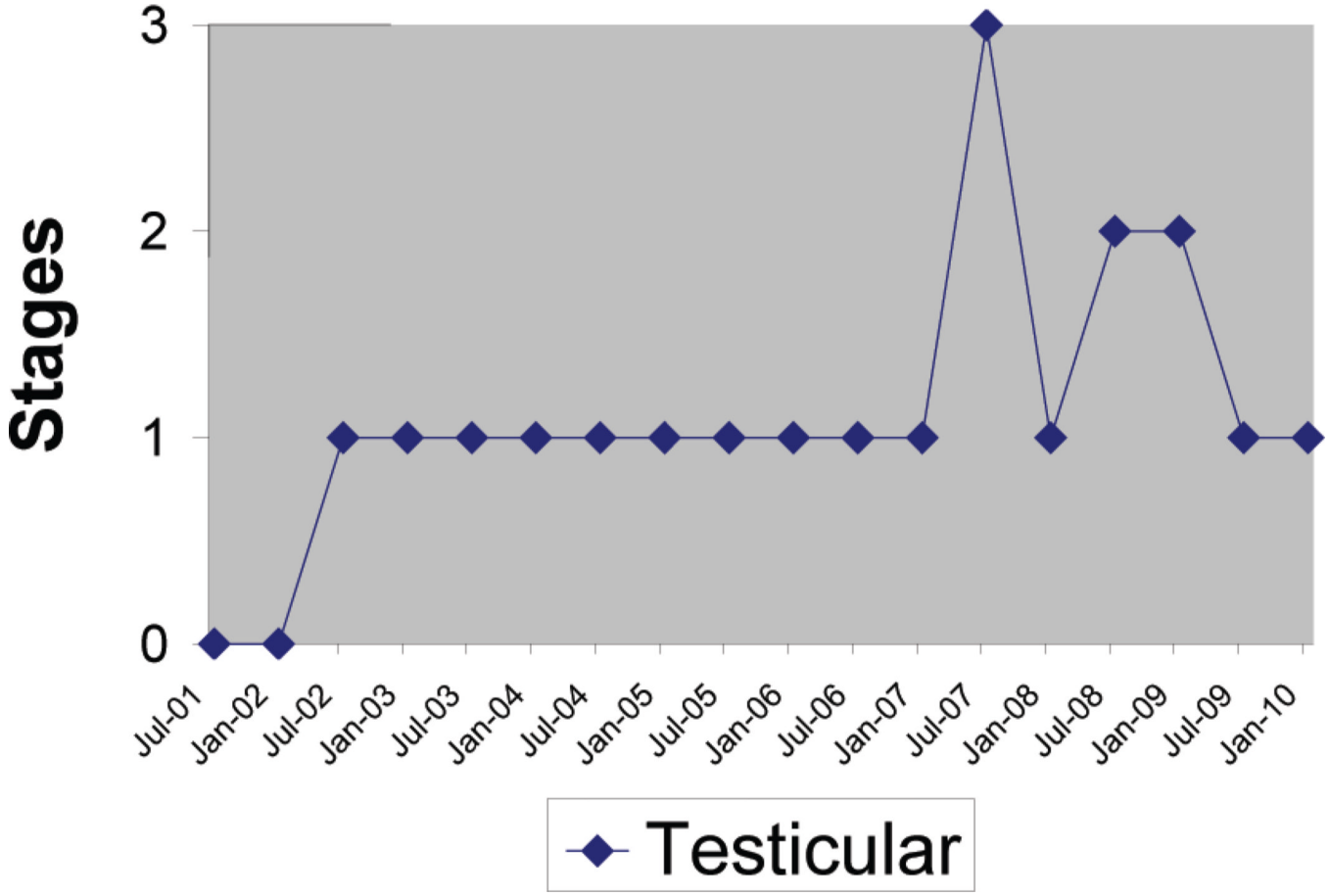

Figure 6.
Testicular cancer forum stage progression

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

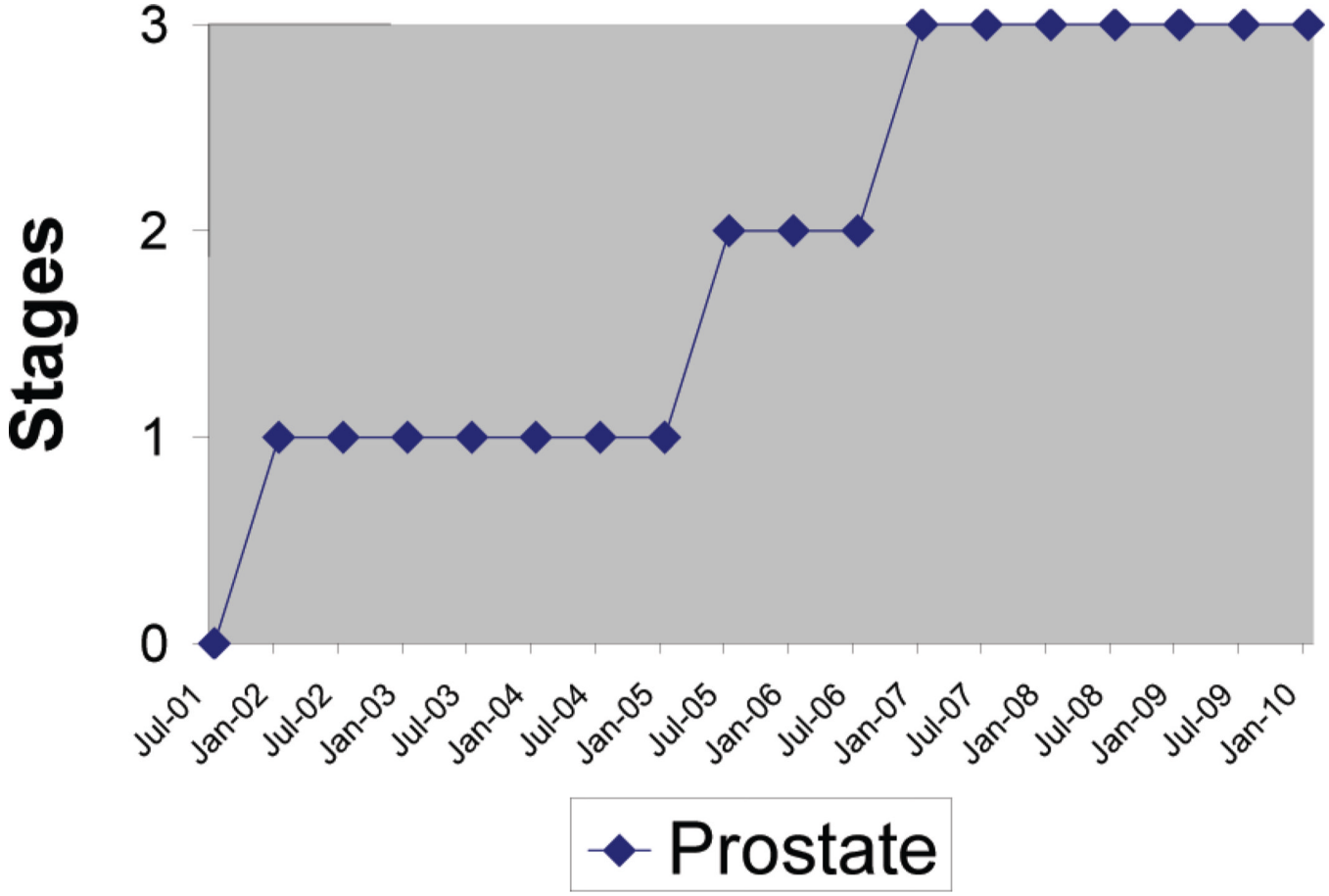

Figure 7.
Prostate cancer forum stage progression

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

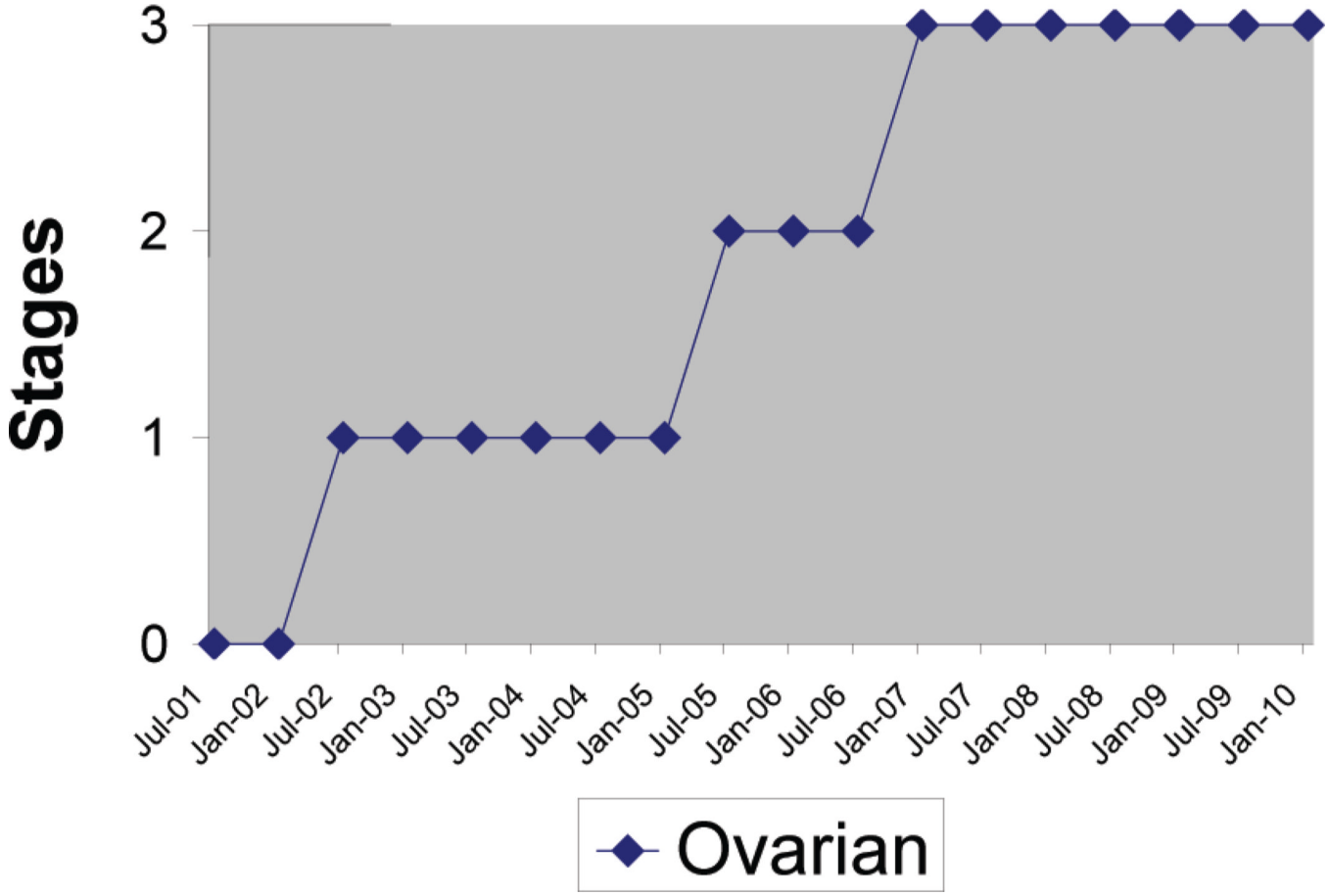

Figure 8.
Ovarian cancer forum stage progression

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

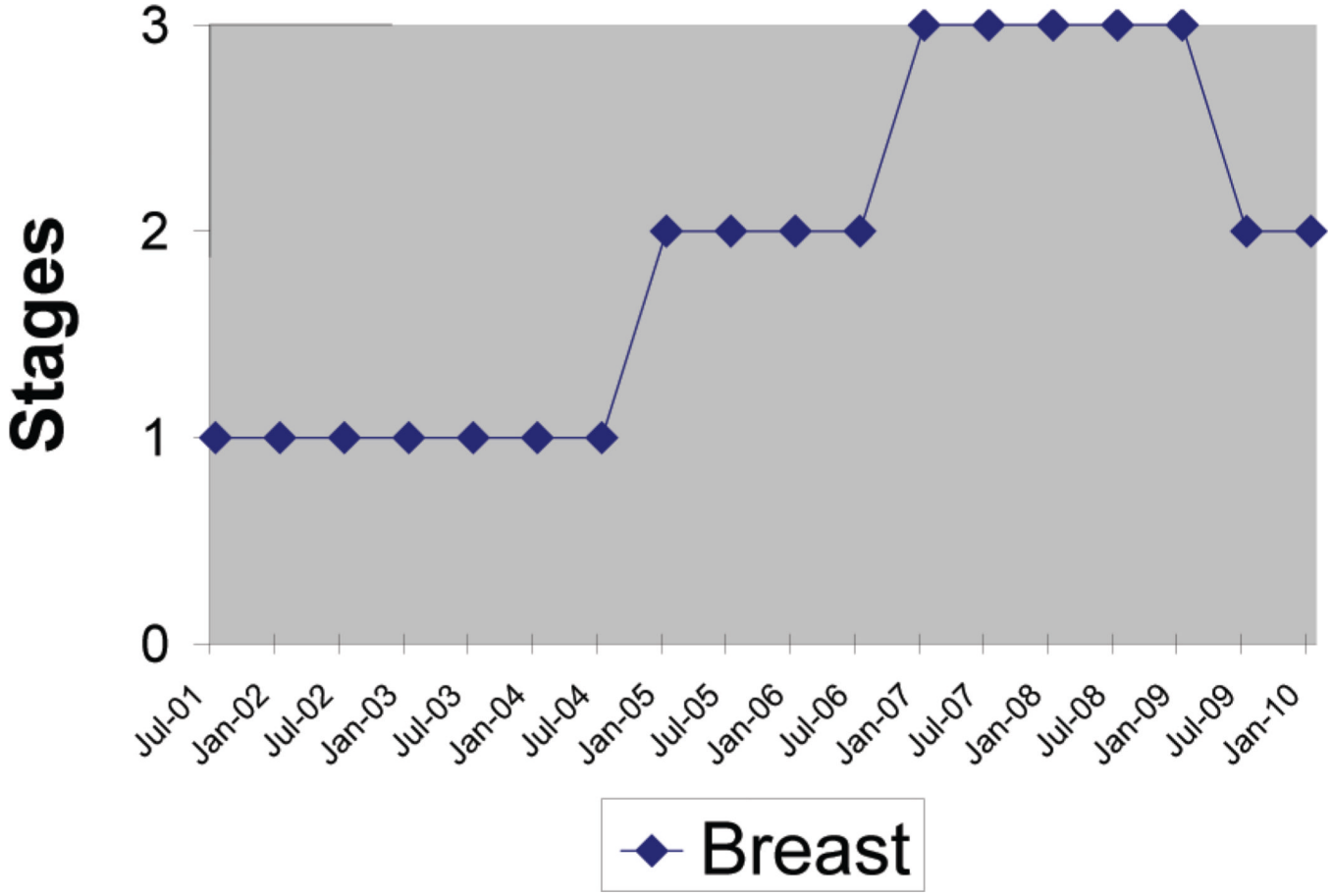

Figure 9.
Breast cancer forum stage progression

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

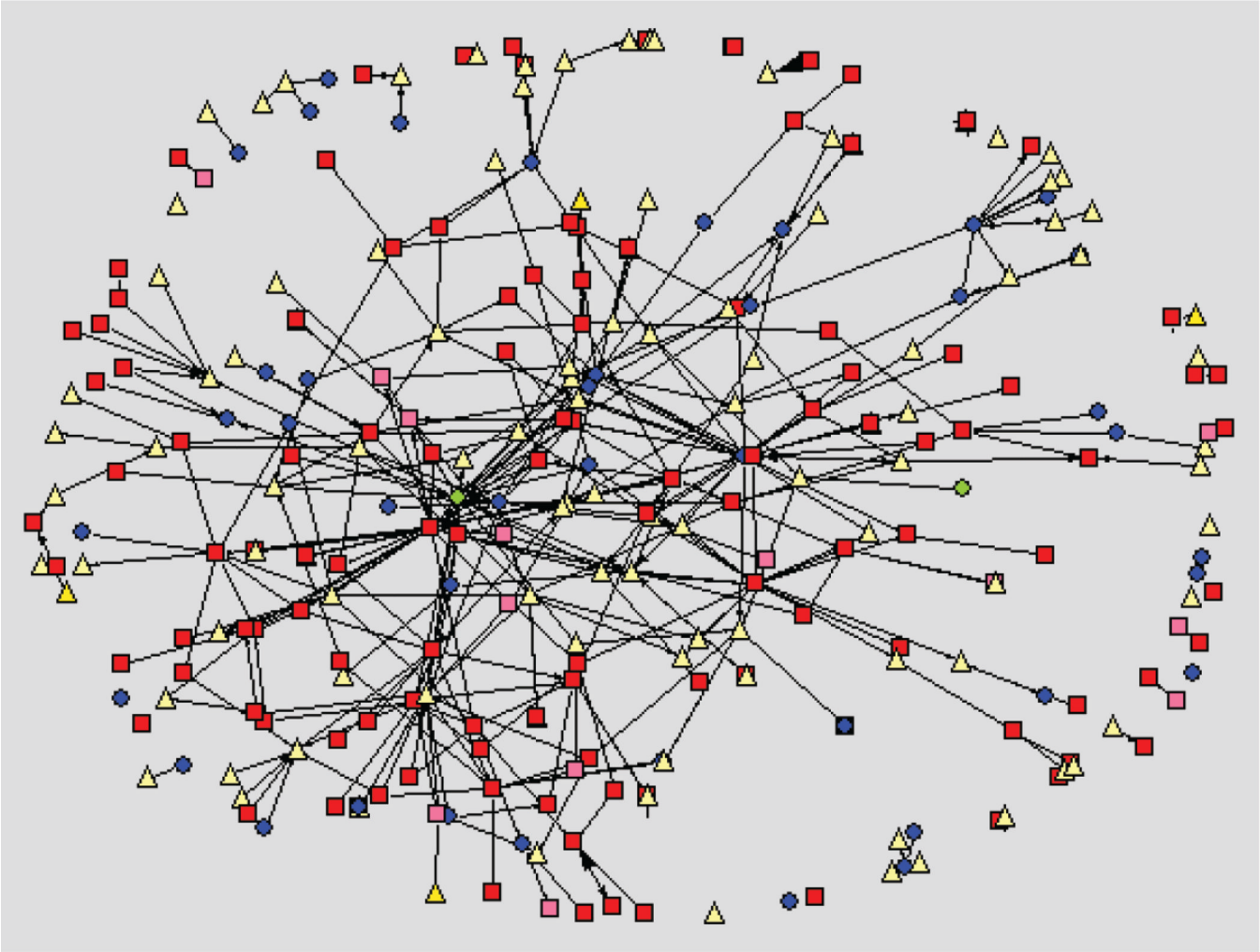

Figure 10.
User connections for the year 2008 within the melanoma forum

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

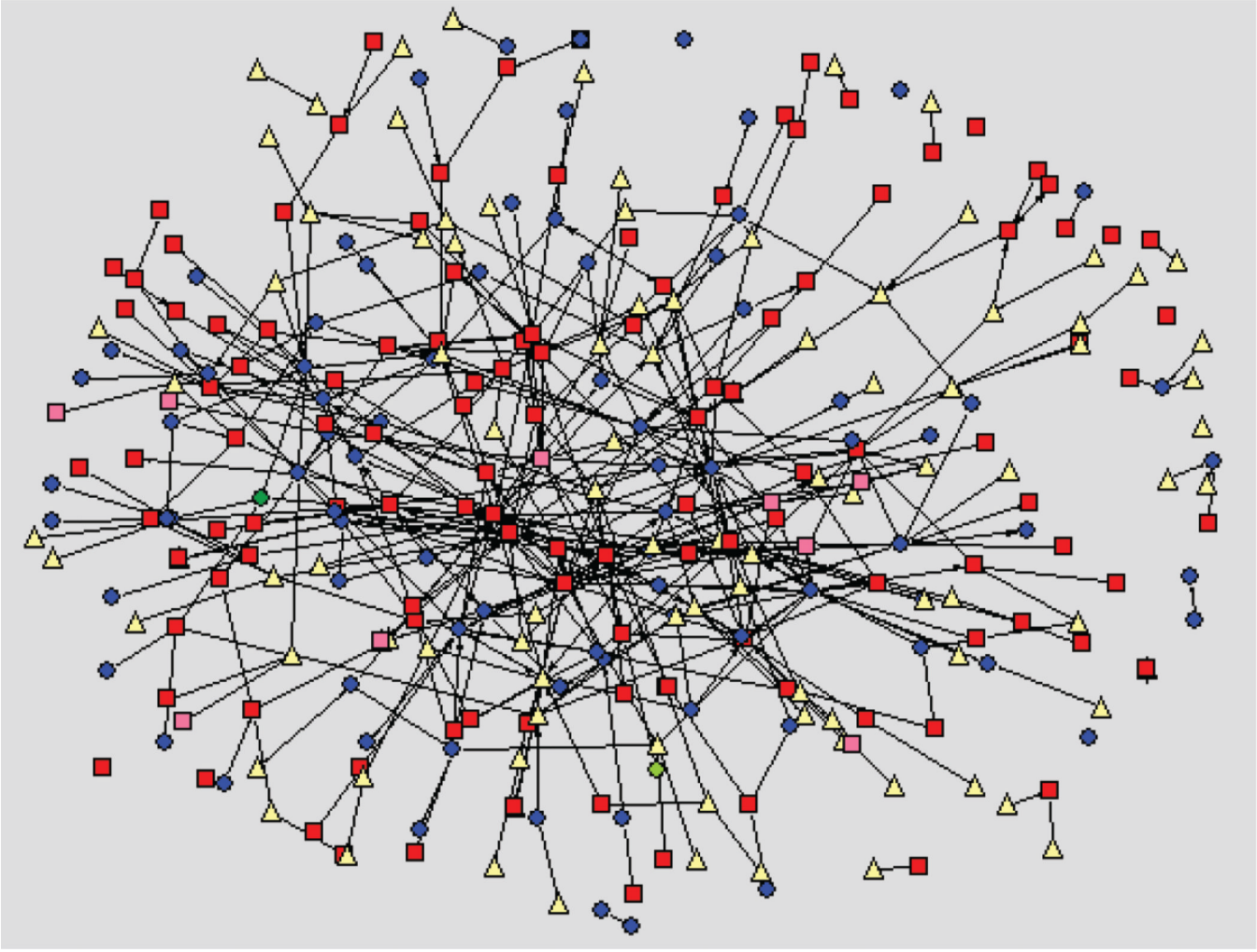

Figure 11.
User connections ions for the year 2009 within the renal cell cancer forum

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

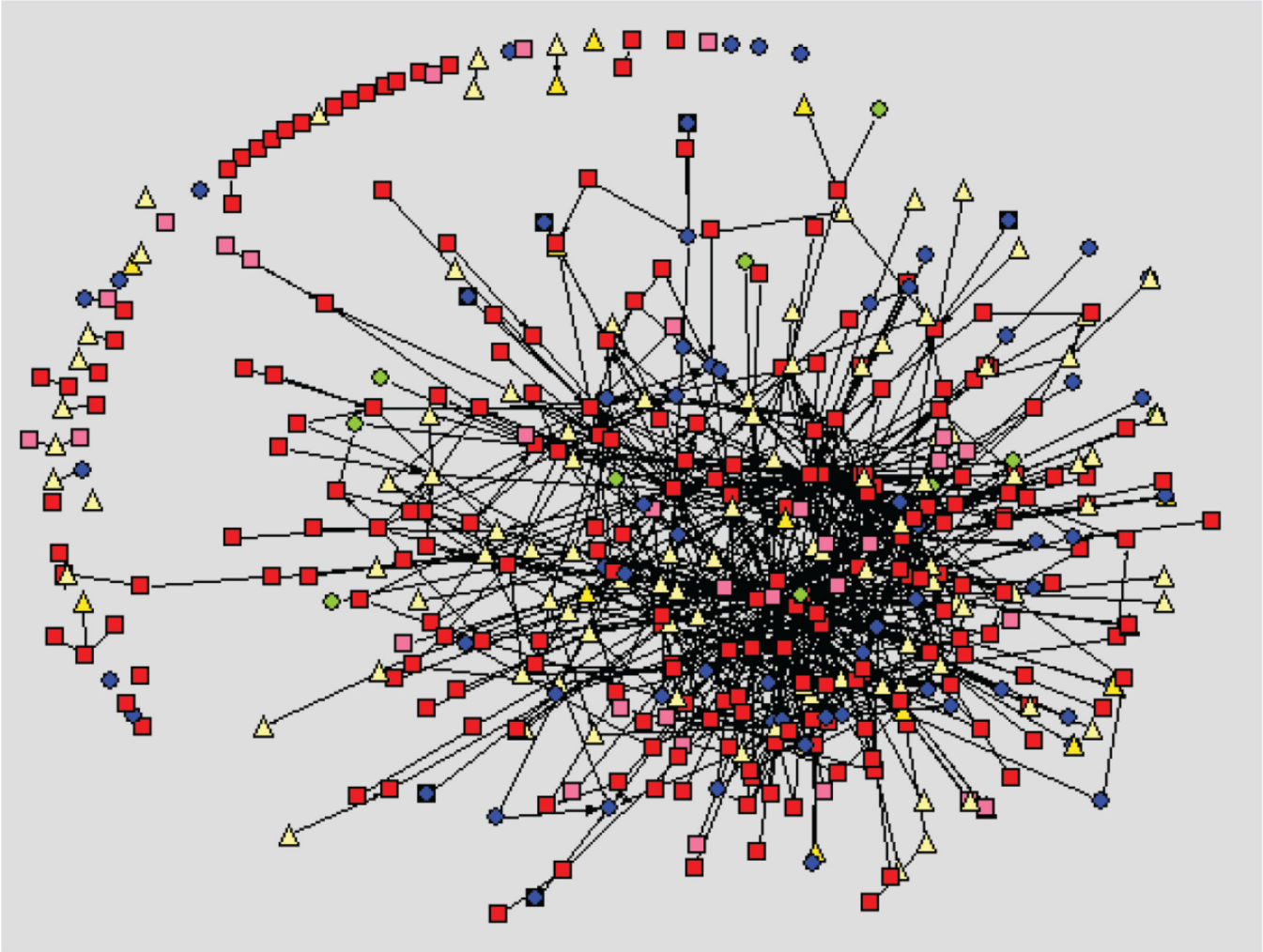

Figure 12.
User connections for the year 2008 within the prostate cancer forum

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

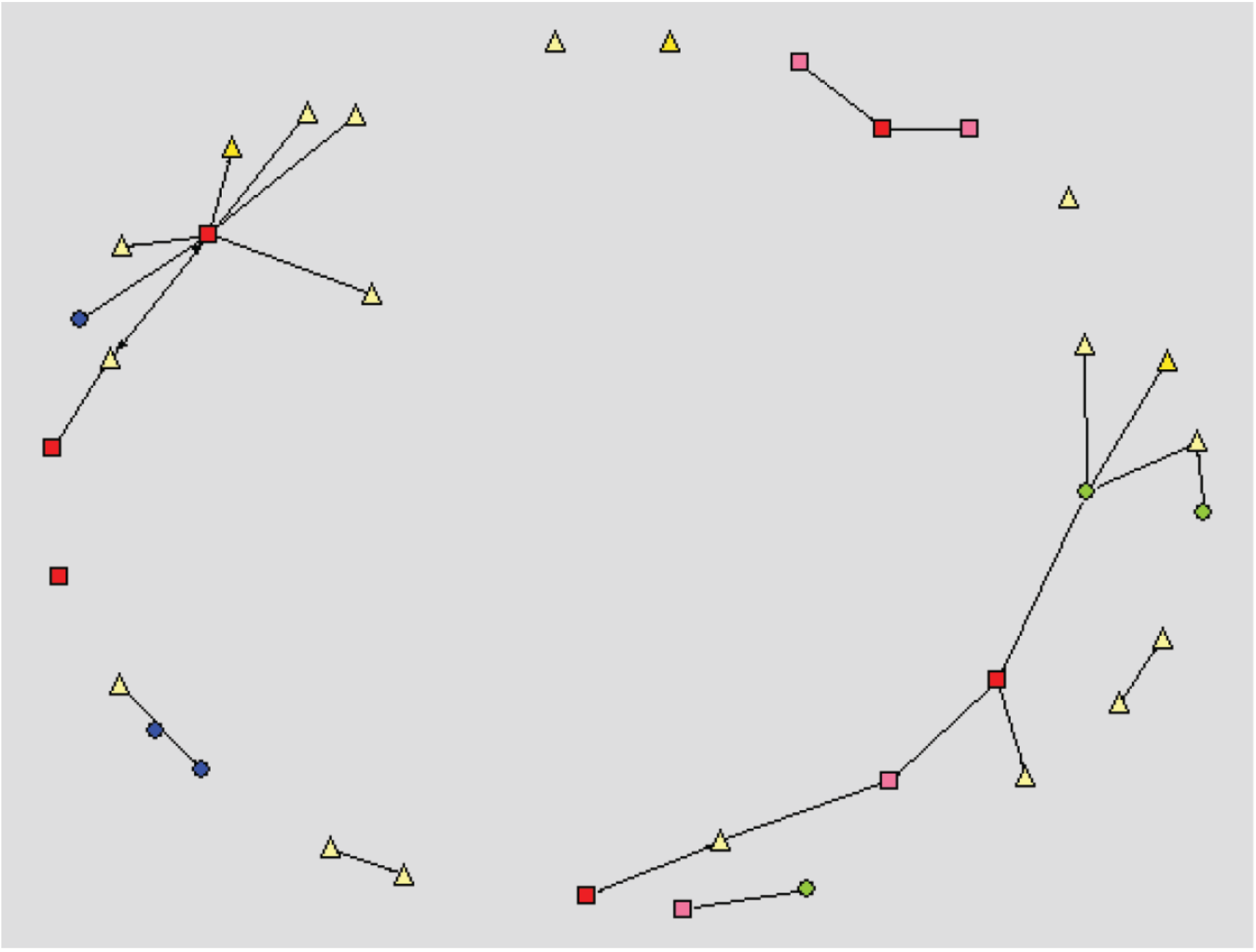

Figure 13.
User connection for the year 2007 within the testicular cancer forum

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

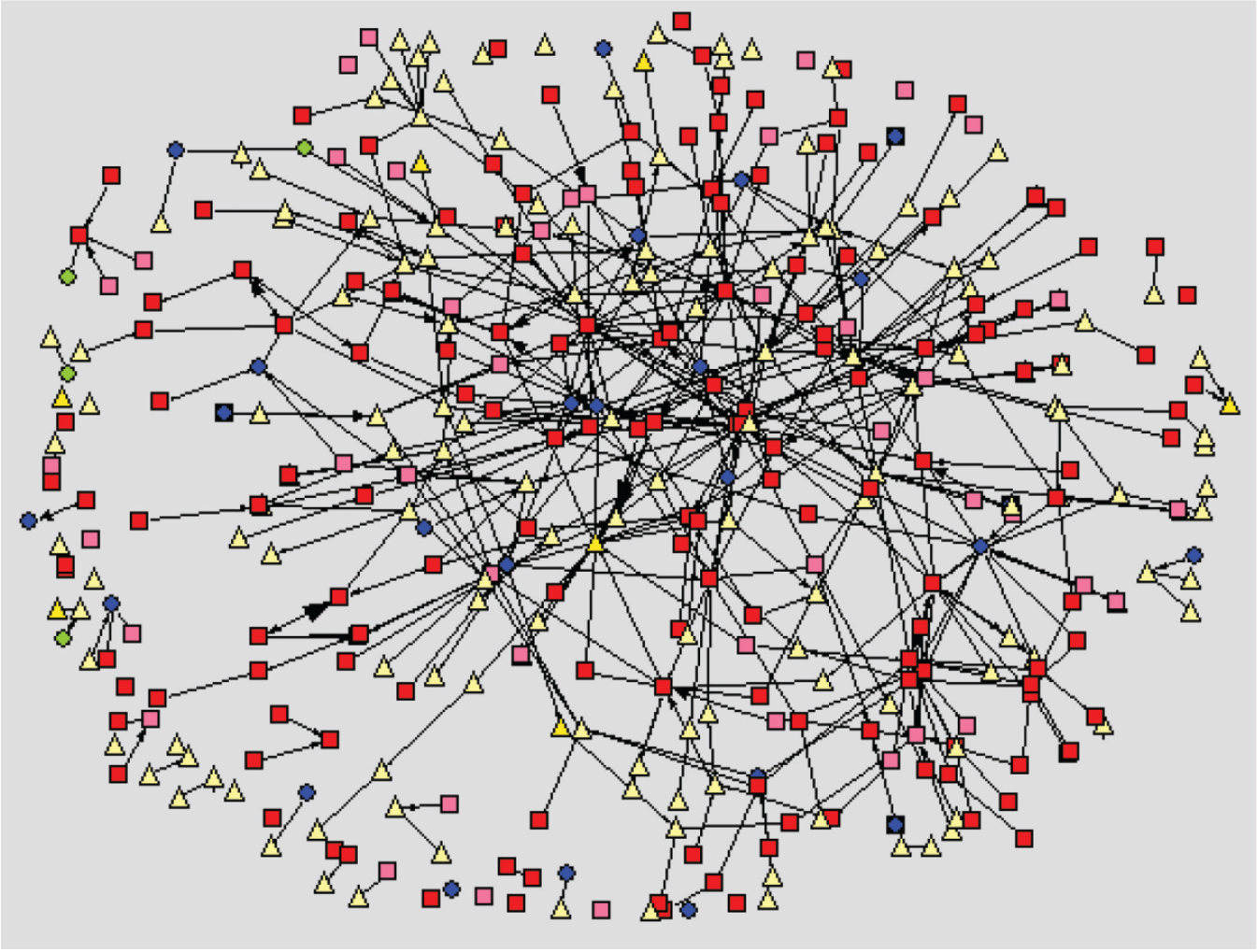

Figure 14.
Users connections for the year 2008 within the ovarian cancer forum

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

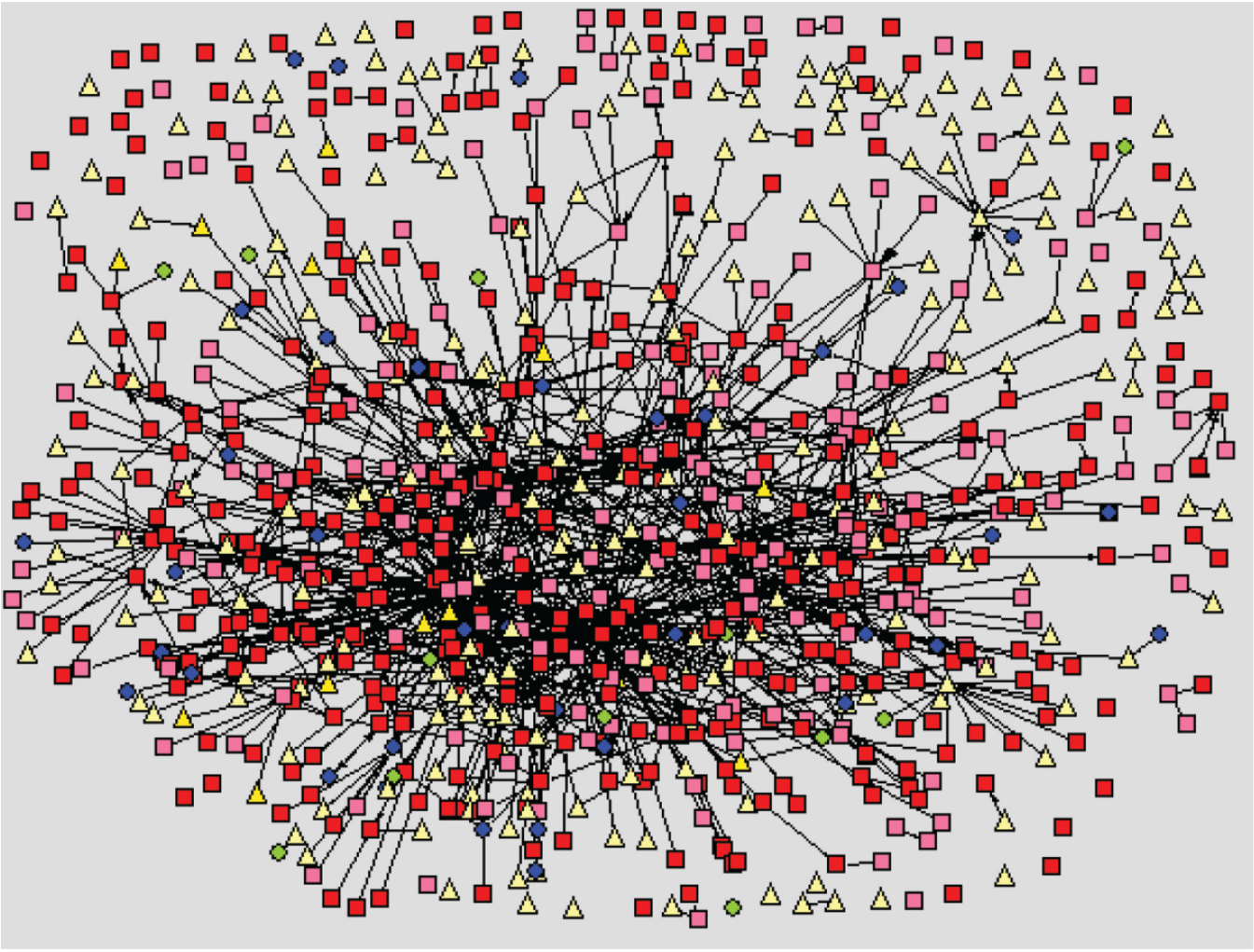

Figure 15.
User connections for the year 2008 within the breast cancer forum

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 1

Number of users stratified by user type

Cancer forum	Melanoma	Renal cell	Prostate	Testicular	Ovarian	Breast
Patient	353	324	611	21	554	1452
Survivor	55	21	82	8	132	589
Caregiver	156	233	231	15	103	169
Doctor	8	6	31	3	15	37
Nurse	1	2	0	0	0	1
Student	0	0	0	0	2	1
Researcher	30	7	47	11	31	66
Unknown	363	311	382	39	515	973
Total	966	904	1384	97	1352	3288
All users						7991

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 2

Number of threads and (posts) created by the different user types

Cancer forum	Mela-noma	Renal cell	Prostate	Test-icular	Ovarian	Breast
Patient	207 (1010)	216 (1322)	567 (3387)	13 (37)	352 (2930)	955 (4489)
Survivor	22 (119)	8 (138)	38 (800)	2 (10)	43 (399)	312 (2175)
Caregiver	80 (450)	193 (1146)	193 (657)	10 (21)	67 (350)	72 (317)
Doctor	3 (122)	2 (20)	4 (411)	0 (6)	3 (42)	9 (242)
Nurse	1 (5)	1 (6)	0 (0)	0 (0)	0 (0)	0 (2)
Student	0 (0)	0 (0)	0 (0)	0 (0)	1 (2)	0 (1)
Researcher	18 (43)	3 (12)	39 (121)	11 (13)	27 (70)	36 (115)
Unknown	273 (981)	251 (927)	350 (1173)	31 (58)	373 (1160)	62 (1963)
Total	604 (2730)	674 (3571)	1191 (6549)	67 (145)	866 (4953)	2009 (9664)
Grand Total						5411 (27612)

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 3

Activity duration for users and threads

Cancer forum	Median User Activity	75% Quartile	Median Thread Activity	75% Quartile
Melanoma	1 [1.0,1.0]	13.1– 1728	12[9.0,17.0]	58.6 –2382
Renal Cell	1 [1.0,2.0]	66.0– 1723	10[8.0–12.0]	38 – 2284
Prostate	1 [1.0,1.0]	36.6–2187	6 [5.0–7.0]	28 – 2364
Testicular	1 [1.0,1.0]	1 – 439	1 [0.0–6.0]	33 –1476
Ovarian	1 [1.0,1.0]	27.6– 2436	12[9.0–14.0]	50 – 2675
Breast	1[1.0,1.0]	14. –2426	7 [6.0–7.0]	30.3– 2993