Supplementary Information for

Activation Forces based Affinity Measure for Analyzing Complex Network

Jun Guo[1], Hanliang Guo[2], Zhanyi Wang[1]

[1]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications

[2]Department of Aerospace and Mechanical Engineering, Viterbi School of Engineering, University of Southern California

Table of Contents

1 Details of the word networks

1.1 Properties of W

To construct a **W**, the parameter *L*, which controls the maximum distance between two words' occurrences that are considered to be a co-occurrence, should be predetermined. Following previous research, we set *L* = 4, 5 and 6, computing three sets of co-occurrence counts and average distances for each pair of words in the vocabulary. Based on these data and the occurrence counts of all the words, we computed three **W**s named here $\mathbf{W}^4$, $\mathbf{W}^5$, and $\mathbf{W}^6$, respectively.

Through the comparison of the elements of the three **W**s, we found that the **W***s* have neither significantly different element values, nor different effects in the experiments. We conclude that the **W**s are not sensitive to the parameter *L* within a certain range (4-6 words). Therefore, we only gave the result from $\mathbf{W}^5$ in our main text for saving space. This is also an implication of the robotics of **W**.

In these **W**s, every particular row or column shows a power-law-like distribution. This means that the coding of **W**s can be very sparse. Namely, we can leave out a large number of *waf*s smaller than a threshold *T* to represent the word network with a few *waf*s that are strong enough to keep the characteristics of the word network. The remaining *waf*s can be considered as the codes for encoding the word network.

In the major experiment, by discarding the *waf*s smaller than *T* = 1.0e-6, the sparseness of **W**s were enhanced at least over 24 times with a cost of losing less than 3.5% sum amount of *waf*s (Table 1). This was a crucial step making the computation of 10,044*10,044 affinities under personal computing environment tractable.

Table 1. The enhancement of sparseness of **W**s

	Num. of nonzero *waf*s (before discarding)	**Num. of nonzero *waf*s (after discarding)**	**Sum amount of *waf*s (before discarding)**	**Sum amount of *waf*s (after discarding)**
$\mathbf{W}^4$	18,497,240	773,011	40.68	39.42
$\mathbf{W}^5$	21,244,909	773,468	39.94	38.61
$\mathbf{W}^6$	23,550,462	764,879	39.10	37.73

In the main text, we have described the heavy tailed distribution of the *waf* values at each node of the network. What does the distribution of *waf* values over the whole network $\mathbf{W}^5$ look like? Fig. 1 gives a clear idea that the distribution is a near-perfect one following a power law with an exponent *k* = 1.1. The sharp descending of the *waf* values guarantees that the spare coding of the **W**s can achieve a high quality, thereby, the word network can be constituted with distinct local structures.

On the other hand, as shown in Fig. 2, the sparse $\mathbf{W}^5$ has the typical feature of a complex network that the link numbers of its nodes distribute following a power law. According to our statistical analysis, the occurrence frequency and the link number of a word in the $\mathbf{W}^5$ are extremely correlated. The correlation coefficients between the frequency and the in-link number, the out-link number, and the total link number are 0.89, 0.91 and 0.92, respectively. These features are completely consistent with the well-known natures of word networks. For the interested readers, the $\mathbf{W}^5$ is available through contact with the corresponding author.

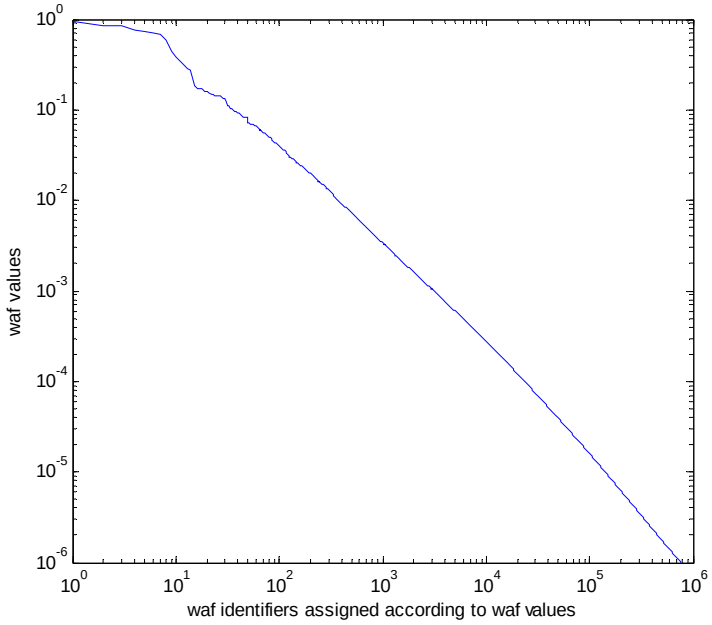

Figure 1 | The distribution of *waf* values over all *waf*s in the sparse $\mathbf{W}^5$. It follows a power law with an exponent k = 1.1. The biggest 10 *waf*s are: *hong-kong* (0.95), *los-angeles* (0.85), *hewlett-packard* (0.84), *gon-na* (0.77), *sri-lanka* (0.74), *per-cent* (0.71), *ulcerative-colitis* (0.69), *saudi-arabia* (0.59), *et-al* (0.44), and *saddam-hussein* (0.38).

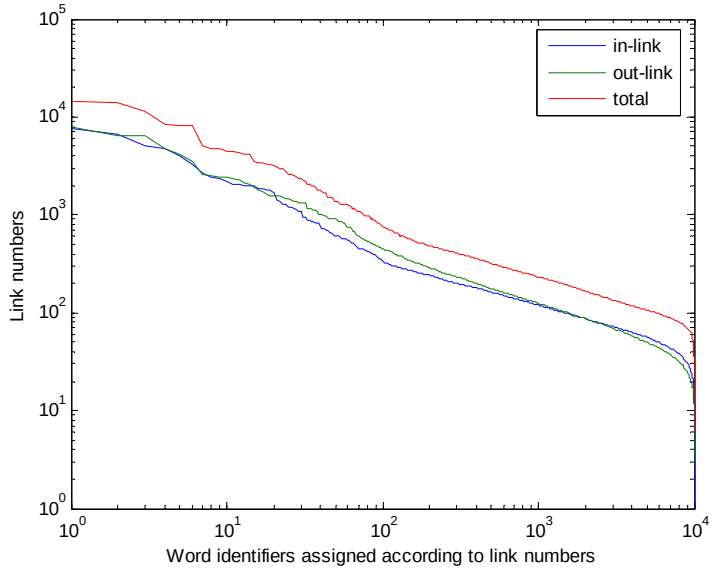

Figure 2 | The distributions of the link numbers of the sparse $\mathbf{W}^5$. The horizontal axis indicates the three sets of word identifiers (No. 1 to No. 10,044) assigned according to in-, out-, and total link numbers, respectively. The figure shows that every distribution of the three follows a power law with the same exponent k = 0.6. The top 5 in-link intensive words are *and* (7590), *the* (6580), *of* (5051), *in* (4800), and *to* (4076); the top 5 out-link intensive words are *the* (7882), *and* (6460), *of* (6352), *a* (4788), and *to* (4097).

1.2 The word affinities based on W

Through a few examples, Fig. 3 shows that by linking a center word and its top 5 neighbours, we easily obtain very meaningful 6-word-clusters, and the local connections of the 6-word-clusters can be striking.

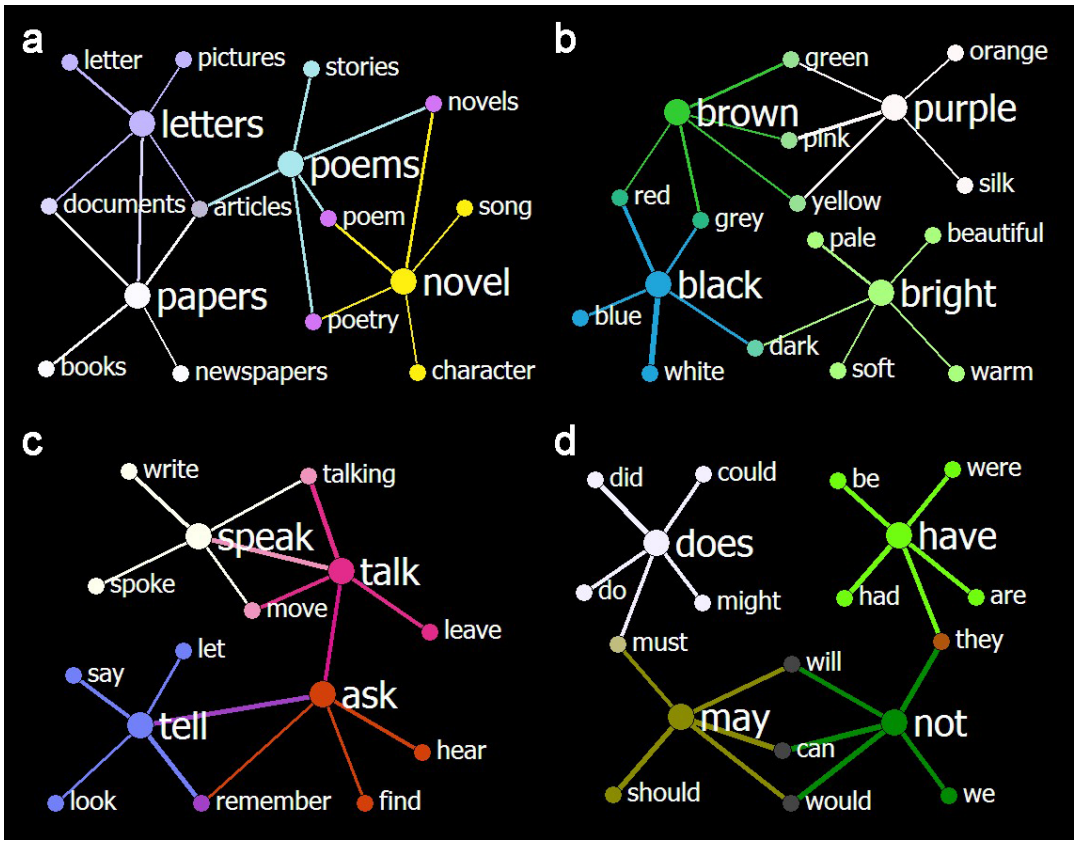

Figure 3 | 6-word clusters and their local connections identified by A^{af}. Center nodes in the clusters are in the bigger size for the eye. The nodes and links belonging to the same cluster are in the same colour except those that are shared by more than one cluster, whose colours are mixed. The thickness of a link represents the affinity between its nodes, ranging from 0.07 (*novel-poetry*) to 0.21 (*have-had*) in this figure. The length of a link means nothing. **a,** Local connections of noun clusters (related to literature). **b,** Adjective-noun clusters (colour). **c,** Verb clusters (talking). **d,** Function word clusters. Besides the plausibility of the clusters and their connections, the strong links between function words are notable.

In addition, Fig. 4 shows a complex sub-network including various hierarchies in the domain of *science and art*. Such a sub-network presents an inherent complex structure of word networks that includes intricate semantic relations.

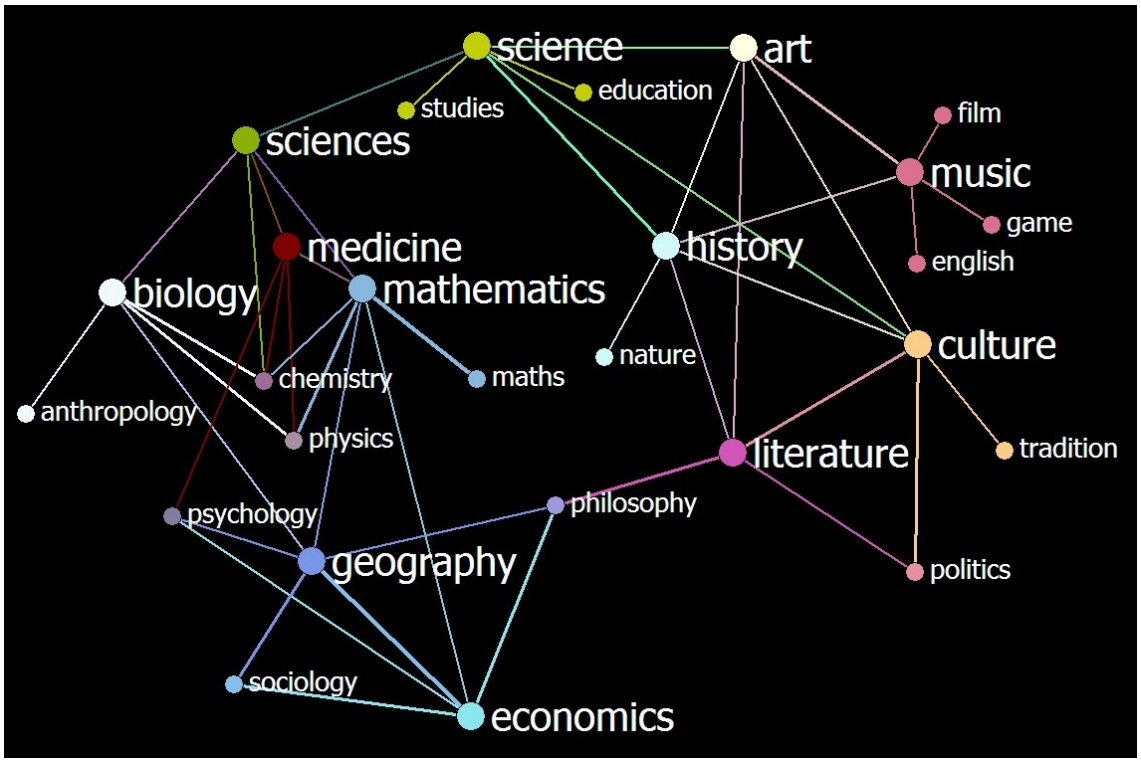

Figure 4 | A complex sub-network composed by 6-word clusters in the domain of *science and art.* The sub-network shows that the affinities between the words are highly sensible, while the hierarchies in the word networks are intricate. In the sub-network, various types of hierarchies are included, for example the vertical hierarchies (*science, sciences, mathematics; art, culture, tradition*), the flat hierarchies (*history, culture, literature; mathematics, geography, economics*), and the hybrid hierarchies (*art, history, music; art, literature, politics*). The affinities range from 0.06 (*medicine-psychology*) to 0.13 (*mathematics-maths*).

1.3 A comparison with human free association

Comparing the neighbouring words in our networks and the associated words in manual free association makes more sense. To this end, targeting 3,269 words in our vocabulary which are overlapped by the words used in the free association data set collected by D. L. Nelson, C. L. McEvoy and T. A. Schreiber, http://w3.usf.edu/FreeAssociation/, we compare their top 3 neighbours in our network and top 3 associates in the free association. The comparison shows that our results are mostly comparable to the ones of the free association. Table 2 presents a small part of the comparison for the target words of common used nouns and verbs in daily life, and a complete comparison is available in Supplementary Data2. From Table 2, we see that for a big proportion of the target words our network and the free association give common answers, such as for *beer*, *wine*, *walk*, *talk* etc. In total this proportion is some 1/4 (798:3,269). For the rest, although both the neighbours and the associates are sensible, they present respective characters. In contrast to the freedom of the associates in categories, the neighbours are exerted more syntactic constraints, leading to that they are much more consistent in parts of speech while they are occasionally loose in semantics (e.g. for *eat*). Notably, our results are merely based on the BNC, a 100 million word corpus, thereby they are naturally characterized by its specific

contexts for each words which are unnecessarily consistent with common sense. For example, in our network, *apple*'s top three neighbours are *microsoft*, *novell* and *ibm*, it suggests that the word is mainly in the contexts of computer industry rather than daily life in the corpus.

Table 2. Top 3 neighbours and associates with target words from our network and free association

Targets	Neighbours of our network	Associates of free association
bread	meat cheese toast	butter dough loaf
butter	cream cheese flour	bread margarine milk
milk	meat cream wine	cow drink honey
drink	drinking coffee sleep	water beer thirst
beer	wine whisky champagne	drink wine drunk
wine	coffee beer champagne	beer drink dine
drunk	asleep alone guilty	alcohol beer drive
drive	driving walk push	car fast way
walk	walking move run	run talk stroll
run	running play move	walk jog fast
sleep	talk drink bed	dream rest awake
talk	speak talking leave	speak listen chatter
leave	stay talk stop	come go arrive
live	lived stay play	die life dead
play	playing played move	fun ball game
move	turn moved talk	leave away stay
ball	shot match straight	bat round throw
throw	pull pick push	ball catch toss
catch	pick throw pull	fish throw ball
fish	animals birds species	water swim sea
water	food light air	drink cool wet
food	material water land	eat drink hunger
eat	talk pick lose	food drink fat
fat	sugar butter diet	skinny thin cat

Except the first one, the targets are iteratively chosen from the previous associates or neighbours.

2 Details of the protein interaction networks

2.1 Properties of P

As mentioned in the main text, **P** represents an un-directional network, which describes the interactions among 4,729 human proteins. The network is characterized by the weighted links of *paf*s. Here we check if the number of links per node still follows a power law distribution like that in the binary protein interaction networks. From Fig. 5 we see that except the tail part, the distribution nearly follows a power law. We suggest that the exception of the tail part is due to the insufficient information on protein interactions we currently know. The highest link number is 221 held by the protein TP53. At another extreme 371 proteins have less than or equal to 3 links, among them 3 proteins have no links above the threshold 1.0e-5.

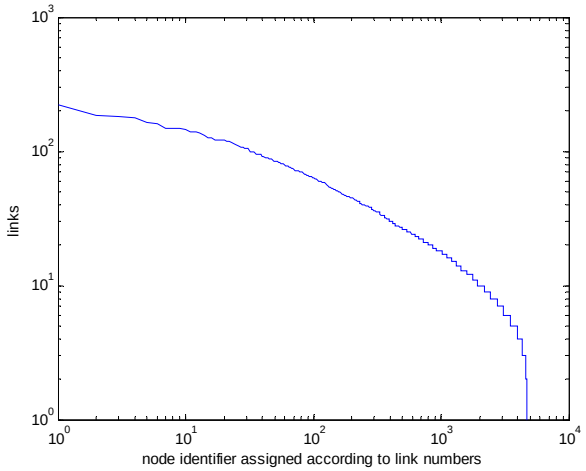

Figure 5 | The distribution of the link numbers over nodes of P.

On the other hand, as shown in Fig.6, the *paf* value distribution over links nearly follows a power law in the period from 1 to 10^4 with the exponent parameter increasing. The highest *paf* value is 0.086 weighting the link between HPS5 and HPS6 while the lowest *paf* value is equal to the threshold 1.0e-5 which is assigned to 67 links. For the interested readers, **P** is available through contact with the corresponding author.

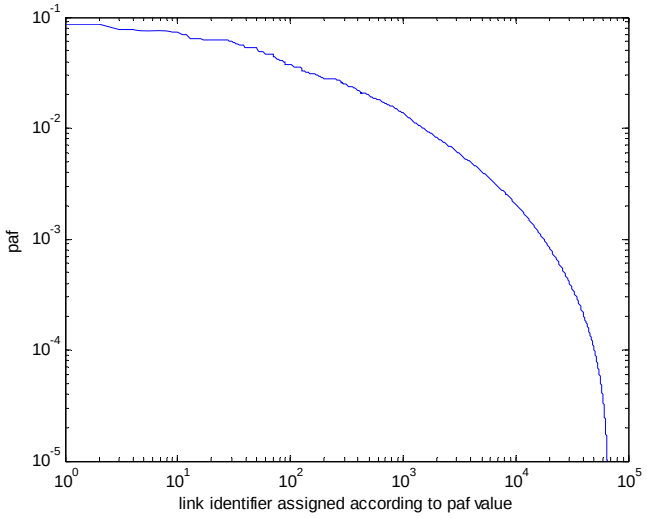

Figure 6 | The distribution of *paf* values over all links.

2.2 Statistics of the affinity networks of CAPs and CAPCs

Currently, HPRD annotates 77 CAPs, 60 out of them are included in **P**, they are:

AXIN2	PPP2R1B	TLR2	BUB1B	EP300	AURKA	MAD1L1	RAD54B	CHEK2	MUTYH
AR	BAX	PTPN12	BRCA2	PMS2	MSH3	PTPRJ	KLF6	STK11	EPHB2
TSG101	PPARG	BARD1	IRF1	IL1B	HRAS	KRAS	SRC	TGFBR2	BRAF
NRAS	PIK3CA	PLA2G2A	APC	PHB	RAD51	RB1	ERBB2	ELAC2	RB1CC1
XRCC3	MSH6	PALLD	BUB1	ATM	PTEN	CASP10	TP53	CDH1	MSH2
MLH1	DCC	EGFR	FAS	FASLG	FGFR3	FGFR4	BRCA1	ZFHX3	CTNNB1

In order to find the CAPCs (CAP closers) which may be significant in cancer study and connect the CAPs into a whole body, we consider two thresholds T_p and T_a, and define that a CAPC have at least T_p CAPs as its neighbours that each have an affinity to it higher than T_a

Table 3 Distribution of the number of links over the CAPs

Number of links	Number of proteins	Gene symbol of proteins
> 19	3	MSH6(27), ATM(22), RAD51(20)
19	2	FASLG, PHB
18	4	BRCA1, MSH2, ERBB2, BRCA2
15	3	PTPN12, PTPRJ, CHEK2
14	2	MLH1, PTEN
12	2	IRF1, BUB1
11	2	AR, AXIN2
10	3	EGFR, BRAF, KLF6
9	6	RB1, KRAS, EPHB2, AURKA, EP300, MUTYH
8	3	SRC, CDH1, RAD54B
7	4	FGFR4, NRAS, APC, MAD1L1
6	1	MSH3
5	4	FAS, PIK3CA, TGFBR2, PMS2
4	5	DCC, HRAS, TP53, PPARG, BUB1B
3	2	BARD1, PPP2R1B
2	8	CTNNB1, FGFR3, BAX, TSG101, CASP10 TLR2, XRCC3, PALLD
1	4	ZFHX3, IL1B, PLA2G2A, ELAC2
0	2	STK11, RB1CC1

By setting T_p = 4 and T_a = 0.03, 82 CAPCs are identified. Incorporating the CAPCs, 58 CAPs form an integral network with affinities higher than 0.03 (not including the ones between CAPCs). In this network, the number of links of a protein is meaningful. The more links a protein has, the more crucial the protein is for the cancer related protein network. Tables 3 and 4 list the distributions of the number of links over the CAPs and CAPCs respectively. From Table 3, we see that the proteins of MSH6, ATM,

RAD51, FASLG, PHB, BRCA1, MSH2, ERBB2 and BRCA2 are notable for their rich links. Actually, these proteins have just been the ones most frequently referred to in cancer study. Table 4 suggests the important roles of the proteins of FLT1, PTMA, RFC1, FANCD2, BLM and TP53BP1 in the connecting of the network. This information may be significant for study of potential cancer related proteins, while the revealing of other CAPCs is informative as well.

Table 4 Distribution of the number of links over the CAPCs

Number of links	Number of proteins	Gene symbol of proteins
8	2	FLT1, PTMA
7	4	RFC1, FANCD2, BLM, TP53BP1,
6	10	JUN, PTPN1, RAP1A, H2AFX, NBN, MDM4 RAD50, ERBB2IP, ATR, TREX1
5	16	EGR1, MUC1, KIT, FER, E2F1, ABL1, HMGA2 PTK2, MRE11A,TCF7L2, CHEK1, BUB3 VDAC1, PPP1R13L, ING4, MDC1
4	50	SELE, ESR1, ERCC5, IFI16, BCR, MIF, CDC25C RELA, NAP1L1, PDGFRB, EPHA2, SOS1, CD28 SP1, CEBPB, PDGFB, WEE1, XRCC1, RRAS2 MST1R, HNF4A, FEN1, SREBF2, SHC1, TUB PTK2B, DAB2, GRB7, TP73, SATB1, MSH4 PLD2, DMC1, DAXX, CCL18, GAB1, WRN TERF2IP, TACC2, SPN, MTA2, CTNNBIP1 NEDD9, CNKSR1, ARHGAP17, AATF, 12257* EFCAB6, CALCOCO1, PBK

* 12257 is the HPRD ID of the protein, which currently has no official gene symbol.

The cancer related protein affinity network is constructed by the CAPs and CAPCs with 445 affinities (links) higher than 0.03. Whereas the distribution of the affinities is skew rather than even (Fig. 7). The average of the affinities is 0.05, while 200 of them are lower than 0.04 and 44 are higher than 0.08. For readers who are interested in the details of the network, the affinities of the whole network are available in Supplementary Data3.

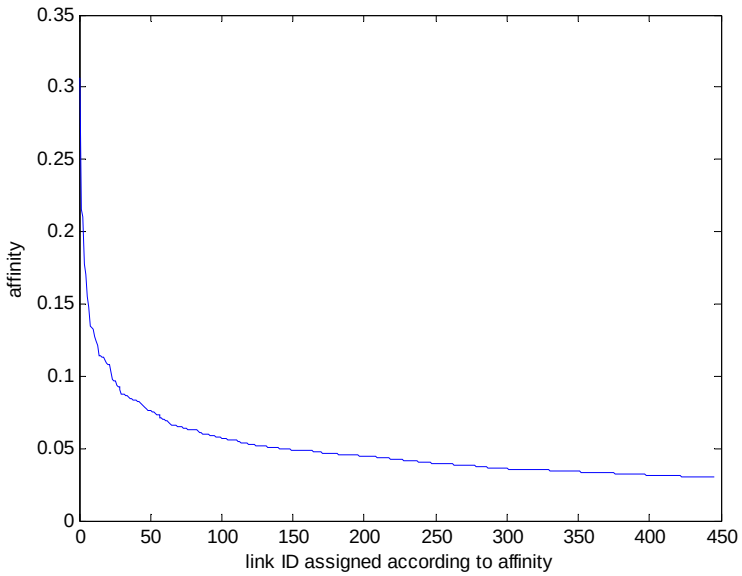

Figure 7 | The distribution of affinities over the links in the cancer related network. The sharp fall at the head and the long flat tail are characteristic.

2.3 Software for the computation of the protein affinity network

The computation of the protein affinity network can simply be implemented according to the definitions of *paf*s and the affinity measure once the basic statistics of PPIs, i.e. occurrence frequencies, co-occurrence frequencies and close distances, are completed. We provide the program for the basic statistics and the program for the computation of the protein affinity network at the web page of http://www.pris.net.cn/download/paf. Readers can also get the programs by sending an email to the corresponding author.