Critiques of applying h-index in YouTube

By Liaonan Xie

The h-index was introduced by Jorge E. Hirsch, a physicist at USCD, in 2005 as a tool for determining theoretical physicists' relative quality based on their productivity and citation impact. Hirsch writes:

"A scientist has index *h* if *h* of his/her N_p papers have at least *h* citations each, and the other $(N_p - h)$ papers have no more than *h* citations each." [Hi05]

This approach was intended to solve the problem of using total number of citations as bibliometric indicator--- a single publication of major influence as one pixel can disproportionately affect the whole picture.

However, criticism argues that h-index may provide misleading information to evaluate a scientist's achievement. One spot is that the h-index does not consider the context of citations; for example, a citation in a paper made in an introduction may not have any direct significant connection to the work. Another is that h-index is bounded by the total number of publications so it indigenously disfavors scientists with short career. Actually it does not provide a significantly more accurate measure of impact than the total number of citations for a given scholar.

Hirsch observed and asserted that h-index is between $0.45\sqrt{N_{citaions}}$ and $0.58\sqrt{N_{citaions}}$.

The rule of thumb for h-index even says:

$$\text{h} = \frac{log2\sqrt{6}}{\pi}\sqrt{N_{citaions}} \approx 0.54\sqrt{N_{citaions}}$$

As shown in the table below, it turns out that the rule provides a highly accurate approximation of *h*-index in most cases for mathematicians. [Yo14]

Medalist	Award year	$N_{citations}$	h	Rule of thumb est.	Confidence interval
T.Gowers	1998	1012	15	17.2	[13, 20]
R.Borcherds	1998	1062	14	17.6	[14, 21]
C.McMullen	1998	1738	25	22.5	[18, 26]
M.Kontsevich	1998	2609	23	27.6	[22, 32]
L.Lafforgue	2002	133	5	6.2	[4, 8]
V.Voevodsky	2002	1382	20	20.0	[16, 23]
G.Perelman	2006	362	8	10.0	[7, 12]
W.Werner	2006	1130	19	18.2	[14, 21]
A.Okounkov	2006	1677	24	22.1	[18, 25]
T.Tao	2006	6730	40	44.3	[38, 51]
C.Ngoˆ	2010	228	9	8.2	[5, 10]
E.Lindenstras	2010	490	12	12.0	[9, 14]
S.Smirnov	2010	521	12	12.3	[9, 15]
C.Villani	2010	2931	25	29.2	[24, 33]

Fields medalists 1998 − 2010

Despite the controversy, however, the influence of h-index is going beyond the scope of academia and producing impact in various fields. One of the interesting applications is in YouTube, , one of the most popular and influential internet media.

Robert Hovden of Cornell University, discussed [Ho13] “the importance in quantitatively evaluating the success of Internet content”, a newly emerged but unexplored field, and used YouTube, as an example to propose potential bibliometrics.

- Hovden defined the h-index for YouTube is the number of videos N that has N×100,000 views or more. In other words, 10^5 video view is analogous to 1 citation in academia. YouTube is acting as the publisher and a particular YouTube channel or user account can be viewed as the author. He also emphasized a single video has only 1 "author" because there is only one uploader.

The following table comes from [Ho13].

Total Views(millions)		**h-index**		**g-index**		**Subscribers(thousands)**	
3280	JustinBieberVEVO	79	Smosh	141	AtlanticVideos	6141	raywilliamjohnson
3175	RihannaVEVO	77	RayWilliamJohnson	130	UltraRecords	6024	nigahiga
2210	AtlanticVideos	70	Nigahiga	128	FueledByRamen	5844	smosh
2184	smosh	69	realannoyingorange	118	smosh	5123	machinima
2177	EminemVEVO	64	UltraRecords	115	realannoyingorange	4706	jennamarbles
2141	RayWilliamJohnson	61	Nqtv	110	barelypolitical	3763	freddiew
2131	LadyGagaVEVO	61	JennaMarbles	109	nigahiga	3222	rihannavevo
1991	UltraRecords	59	MondoMedia	104	linkinparktv	3123	collegehumor
1834	shakiraVEVO	58	AtlanticVideos	101	kontor	2982	shanedawsontv
1726	FueledByRamen	58	Fred	99	nqtv	2920	fpsrussia
1668	beyonceVEVO	57	huluDotCom	97	Fred	2861	epicmealtime
1608	officialpsy	56	barelypolitical	96	SpinninRec	2715	pewdiepie
1553	barelypolitical	55	Muyap	96	huluDotCom	2690	bluexephos
1498	hollywoodrecords	55	Freddiew	94	MondoMedia	2573	realannoyingorange
1487	realannoyingorange	54	Kontor	93	RovioMobile	2515	thelonelyisland
1445	BlackEyedPeasVEVO	54	BritainsGotTalent09	93	JennaMarbles	2499	tobuscus
1439	ChrisBrownVEVO	54	Boyceavenue	92	BritainsGotTalent09	2500	kevjumba
1429	muyap	50	Machinima	92	TheOfficialSkrillex	2460	werevertumorro
1423	machinima	48	FueledByRamen	92	davidguetta	2417	riotgamesinc
1421	JenniferLopezVEVO	48	TheXFactorUK	90	Flowgo	2360	michellephan
1411	kontor	47	beyonceVEVO	88	sment	2333	roosterteeth
1384	PitbullVEVO	47	ShaneDawsonTV	88	RayWilliamJohnson	2325	onedirectionvevo
1376	KatyPerryVEVO	46	collegehumor	86	warnerbrosrecords	2292	justinbiebervevo
1354	MondoMedia	46	warnerbrosrecords	84	daneboe	2253	sxephil
1336	nigahiga	44	SpinninRec	83	thelonelyisland	2143	barelypolitical

Based upon the 50 most subscribed channels, Hovden showed in [Ho13] the Pearson correlation coefficients of the YouTube h-index, g-index, and total views to a channel's subscribers, are 0.68, 0.47, and 0.38, with p-values of 1.8×10^{-8}, 4.0×10^{-4}, and 5.0×10^{-3} respectively. "These values indicate that the h-index has the strongest correlation with the number of subscribers when considering top YouTube channels." Based on the correlation above seemingly h-index is a potential tool to measure the impact of YouTube channels.

However, things in YouTube are radically different from things in academia. There are a few questions about using h-index in YouTube.

Top Rankings for Different YouTube Channel Types

Comedians			Musicians		
	h-index	total-views		*h*-index	total-views
'smosh'	79	2153	'UltraRecords'	64	1990
'RayWilliamJohnson'	77	2140	'AtlanticVideos'	58	2209
'nigahiga'	70	1304	'boyceavenue'	54	808
'realannoyingorange'	69	1486	'kontor'	54	1410
'nqtv'	61	1021	'FueledByRamen'	48	1725
'Fred'	58	949	'beyonceVEVO'	47	1667
'collegehumor'	46	1136	'UKFDubstep'	42	920
'AdamThomasMoran'	44	385	'RihannaVEVO'	41	3172
'TheEllenShow'	41	1052	'shakiraVEVO'	39	1833
'werevertumorro'	40	732	'linkinparktv'	37	1096
Gurus			**Reporters**		
	h-index	total-views		*h*-index	total-views
'FPSRussia'	40	490	'AssociatedPress'	31	609
'MichellePhan'	38	626	'Matroix'	19	262
'kipkay'	33	378	'ABCNews'	18	290
'Howcast'	25	524	'www16barsde'	16	145
'expertvillage'	24	517	'JuliensBlog'	15	116
'bubzbeauty'	21	256	'TMZ'	14	127
'HouseholdHacker'	20	201	'IshatOnU'	14	110
'CaptainSparklez'	19	591	'CTFxC'	12	186
'TobyGames'	19	423	'FUNKER530'	12	105
'dope2111'	18	143	'scoutthedoggie'	12	109

From the table above [Ho13], it is not hard to notice, "UKFDubstep" has only 920 * 10^5 total-views but has 42 h-index while "RihannaVEVO" has 3172* 10^5 total-views but only 41 h-index. This difference clearly obeys the $0.45\sqrt{N_{citaions}}$ and $0.58\sqrt{N_{citaions}}$ bounds and the rule of thumb. The following table shows comprehensively how the lower bound, upper bound and rule of thumb go wrong in YouTube world.

Comedians	h-index	lower bound	upper bound	rule of thumb	total-views
smosh'	79	20.88019396	26.91225	25.05623276	2153
RayWilliamJohnson'	77	20.81706031	26.83087773	24.98047237	2140
nigahiga'	70	16.24992308	20.9443453	19.49990769	1304
realannoyingorange'	69	17.34690174	22.35822891	20.81628209	1486
nqtv'	61	14.37889078	18.53279256	17.25466893	1021
Fred'	58	13.86262962	17.86738929	16.63515554	949
collegehumor'	46	15.16706959	19.54866747	18.20048351	1136
AdamThomasMoran'	44	8.829637592	11.38042178	10.59556511	385
TheEllenShow'	41	14.59554727	18.8120387	17.51465672	1052
'werevertumorro'	40	12.17497433	15.69218914	14.6099692	732

Muscians	h-index	lower bound	upper bound	rule of thumb	total-views
'UltraRecords'	64	20.07423722	25.87346131	24.08908467	1990
'AtlanticVideos'	58	21.15	27.26	25.38	2209
'boyceavenue'	54	12.79140336	16.48669767	15.34968404	808
'kontor'	54	16.89748502	21.77898069	20.27698202	1410
'FueledByRamen'	48	18.68990369	24.0892092	22.42788443	1725
'beyonceVEVO'	47	18.3730101	23.68076857	22.04761212	1667
'UKFDubstep'	42	13.6491758	17.59227103	16.37901096	920
'RihannaVEVO'	41	25.34423011	32.66589659	30.41307614	3172
'shakiraVEVO'	39	19.26609717	24.83185857	23.1193166	1833
'linkinparktv'	37	14.89765082	19.20141661	17.87718099	1096

Gurus	h-index	lower bound	upper bound	rule of thumb	total-views
'FPSRussia'	40	9.96117463	12.8388473	11.95340956	490

'MichellePhan'	38	11.2589964	14.51159536	13.51079568	626
'kipkay'	33	8.748999943	11.27648882	10.49879993	378
'Howcast'	25	10.30097083	13.27680685	12.36116499	524
'expertvillage'	24	10.2319353	13.18782772	12.27832236	517
'bubzbeauty'	21	7.2	9.28	8.64	256
'HouseholdHacker'	20	6.379851095	8.22291919	7.655821315	201
'CaptainSparklez'	19	10.9397212	14.10008511	13.12766544	591
'TobyGames'	19	9.255133711	11.928839	11.10616045	423
'dope2111'	18	5.381217334	6.935791231	6.457460801	143

Reporters	h-index	lower bound	upper bound	rule of thumb	total-views
'AssociatedPress'	31	11.10506641	14.31319671	13.32607969	609
'Matroix'	19	7.283886325	9.388120153	8.74066359	262
'ABCNews'	18	7.663223865	9.877044092	9.195868638	290
'www16barsde'	16	5.41871756	6.984124856	6.502461073	145
'JuliensBlog'	15	4.846648326	6.246791176	5.815977992	116
'TMZ'	14	5.071242451	6.536268048	6.085490942	127
'IshatOnU'	14	4.719639817	6.083091319	5.66356778	110
'CTFxC'	12	6.137181764	7.910145384	7.364618116	186
'FUNKER530'	12	4.611127845	5.943231444	5.533353414	105
'scoutthedoggie'	12	4.698137929	6.055377775	5.637765515	109

The actual h-index is significantly larger than the "expected h-index". One of the reasons behind this difference may be related to the number 100,000.

From Hovden's definition, the number 10^5 is striking. It seems to be a random number.

He also noticed the choice of 100,000 could be a problem and tried to verify it. [Hi05] He explained that this magnitude produces h-index values of top YouTubers most consistent with the top academics. In Hirsche's paper, the mean and median h-index of Nobel Prize winning physicists (years 1975-2005) are 41 and 35 respectively. Using 100,000 views, the top 25 YouTube h-indexes have a mean and median of 56.7 and 55. However, this claim is problematic.

First of all, 41 and 35 is different from 56.7 and 55. In terms of h- index, the word "consistent" is ambiguous. Second, as he himself mentioned in his paper, in academic publications, "the h-index

is criticized for its poor ability in comparing scholars from different fields with different citation behavior.” [Ho13]. Based on Professor Alexander Yong's paper [Yo14], the h-index of noble mathematicians, Fields medalists 1998 − 2010, is substantially below the 41 and 35 line, as shown in the previous table. Therefore, if it is not appropriate to compare scholars from different fields based on h-index, certainly the h-index of YouTubers should be used to compare with that of top academics. Theoretically, the method of choosing 100,000 is unproved.

In practice, choosing the number of views analogous to 1 citation is difficult because the number cannot be deterministic.

In his paper [Ho13], Hovden only discussed the Top 25 YouTubers. Even if assumed 100,000 is the number, theoretically correct, and it works for the calculation and comparison of Top 25 YouTubers' h-index, the number may not satisfy the need to calculate and compare the h-index of strictly academic YouTube videos or small community or family YouTube videos. For example, the h-index based on 100,000 of two science and technology channel A and B could both be 0, but all 40 videos of channel A have at least 10,000 views each while only 10 videos of channel B have at least 10,000 views.

Furthermore, all of the above discussion is based on the assumption that the view is similar to citation; 1 number of view means 1 person/IP address visited the video and 1 citation means 1 scholar cited the work in his/her paper. Unfortunately this assumption is invalid. According to Ted Hamilton, a product manager for YouTube Analytics, the department responsible for managing the counting of YouTube views, “view is their currency” so they try to eliminate the so call “counterfeit views”. In order to do so, YouTube developed the following algorithm.

```
if (view count<= 300)        view count= view count +1;

else                         go to program X;
```

The algorithm says if the view count is smaller or equal to 300, view count will be incremented by 1. After view count reaches 301, a program called X will be invoked, where X decides if view count should be incremented and by how much view count should be incremented.

While detail about X remains a business secret, some information about X is known to the public. One is that N visits from the same IP address to one video do not generate N view counts. The number of view generated by this N visits from same IP address to one video depends on many other factors such as patterns of visits (for example, the length of the video is played) and additional visits from the same IP address to related videos within certain period of time.

Despite the complexity behind the number of views, it is clear that H views cannot be used as analogous to 1 citation, unless number H is smaller or equal to 300. It is undesirable that 1 citation in Einstein's paper is weighted differently from 1 citation in other physicists' paper, so the artificial number of views cannot meet the most fundamental requirement of the h-index.

In conclusion, applying the h-index to YouTube is a desirable idea, because we will have a scientific tool to quantitatively measure this representative of a form of media. Hovden created a captivating skeleton but we still need to solve problems discussed above. Once the h-index can be successfully applied to YouTube, similar measurement for Facebook, Twitter, or LinkedIn might create many unexpected social impact and reshape the way people see this world.

REFERENCES

[Yo14] A.Yong, Critique of Hirsch's citation index: a combinatorial Fermi problem, Notices of the American Mathematical Society, 2014 October 13.

[Hi05] J. E. Hirsch, An index to quantify an individual's scientific research output, Proc Natl Acad Sci USA. 2005 November 15; 102(46): 16569–16572.

[Ho13] R. Hovden, "Bibliometrics for Internet Media: applying the h-index to YouTube. Journal of the American Society for Information Science & Technology. 2013; 64(11): 2326-2331.